

Shahram Latifi *Editor*

ITNG 2021 18th International Conference on Information Technology-New Generations

Advances in Intelligent Systems and Computing

Volume 1346

Series editors

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

****Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink****

Advisory Board

Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

e-mail: nikhil@isical.ac.in

Members

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing, Universidad Central de Las Villas, Santa Clara, Cuba

e-mail: rbellop@uclv.edu.cu

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

escorchado@usal.es

Hani Hagrais, School of Computer Science & Electronic Engineering, University of Essex, Colchester, UK

e-mail: hani@essex.ac.uk

László T. Kóczy, Department of Information Technology, Faculty of Engineering Sciences, Győr, Hungary

e-mail: koczy@sze.hu

Vladik Kreinovich, Department of Computer Science, University of Texas at El Paso, El Paso, TX, USA

e-mail: vladik@utep.edu

Chin-Teng Lin, Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan

e-mail: ctlin@mail.nctu.edu.tw

Jie Lu, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, Australia

e-mail: Jie.Lu@uts.edu.au

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute of Technology, Tijuana, Mexico

e-mail: epmelin@hafsamx.org

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro, Rio de Janeiro, Brazil

e-mail: nadia@eng.uerj.br

Ngoc Thanh Nguyen, Faculty of Computer Science and Management, Wrocław University of Technology, Wrocław, Poland

e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

Jun Wang, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

e-mail: jwang@mae.cuhk.edu.hk

More information about this series at <http://www.springer.com/series/11156>

Shahram Latifi
Editor

ITNG 2021 18th International Conference on Information Technology-New Generations

 Springer

Editor
Shahram Latifi
Department of Electrical and Computer
Engineering
University of Nevada
Las Vegas, NV, USA

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-030-70415-5 ISBN 978-3-030-70416-2 (eBook)
<https://doi.org/10.1007/978-3-030-70416-2>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I AI and Robotics

1	Conceptualisation of Breast Cancer Domain Using Ontology	3
	Reshmy Krishnan, P. C. Sherimon, and Menila James	
2	Traffic Light Control and Machine Learning: A Systematic Mapping Review	11
	Dimitrius F. Borges, Edmilson M. Moreira, Adler D. de Souza, and João Paulo R. Leite	
3	Human-in-the-Loop Flight Training of a Quadcopter for Autonomous Systems	19
	Luke Rogers and Alex Redei	
4	COVID-19: The Importance of Artificial Intelligence and Digital Health During a Pandemic	27
	Maximilian Espuny, José S. da Motta Reis, Gabriel M. Monteiro Diogo, Thalita L. Reis Campos, Vitor H. de Mello Santos, Ana C. Ferreira Costa, Gildarcio S. Gonçalves, Paulo M. Tasinaffo, Luiz A. Vieira Dias, Adilson M. da Cunha, Nilo A. de Souza Sampaio, Andréia M. Rodrigues, and Otávio J. de Oliveira	
5	CropWaterNeed: A Machine Learning Approach for Smart Agriculture	33
	Malek Fredj, Rima Grati, and Khoulood Boukadi	
6	Machine Learning: Towards an Unified Classification Criteria	39
	Clara Burbano, David Reveló, Julio Mejía, and Daniel Soto	

Part II Cybersecurity I

7	Classification and Update Proposal for Modern Computer Worms, Based on Obfuscation	49
	Hernaldo Salazar and Cristian Barria	
8	Conceptual Model of Security Variables in Wi-Fi Wireless Networks: Review	59
	Lorena Galeazzi, Cristian Barría, and Julio Hurtado	
9	Cybersecurity Analysis in Nodes that Work on the DICOM Protocol, a Case Study	69
	David Cordero and Cristian Barría	
10	Hybrid Security Risk Assessment Model	77
	Robert Banks, Jim Jones, Noha Hazzazi, Pete Garcia, and Russell Zimmermann	

11	Enriching Financial Software Requirements Concerning Privacy and Security Aspects: A Semiotics Based Approach	85
	Leonardo Manoel Mendes, Ferruccio de Franco Rosa, and Rodrigo Bonacin	
12	Efficient Design of Underwater Acoustic Sensor Networks Communication for Delay Sensitive Applications over Multi-hop	91
	Ahmed Al Guqhaiman, Oluwatobi Akanbi, Amer Aljaedi, and C. Edward Chow	
Part III Cybersecurity II		
13	Parallelized C++ Implementation of a Merkle Tree	107
	Andrew Flangas, Autumn Cuellar, Michael Reyes, and Frederick C. Harris Jr.	
14	A Study on Ontologies of Vulnerabilities and Attacks on VLAN	115
	Marcio Silva Cruz, Ferruccio de Franco Rosa, and Mario Jino	
15	Towards a Symmetric Crypto Algorithm: The HAJ	121
	Daniel Alarcón-Narváez and Fausto A. Jacques García	
16	A Comparative Study Between Two Numerical Methods for Symmetric Cryptography Uses and Applications	127
	Alba Nidia Martínez-Martínez and Fausto A. Jacques García	
17	Speed Up Over the Rainbow	131
	Nir Drucker and Shay Gueron	
18	Extending a Hybrid Security Risk Assessment Model with CWSS	137
	Robert Banks, Jim Jones, Noha Hazzazi, Pete Garcia, and Russell Zimmermann	
Part IV E-Health		
19	Identifying and Prioritizing Applications of Internet of Things in the Supply Chain of Distribution and Sale of Health Care Products in Iran	147
	Niloofer AminiKalibar and Fatemeh Saghafi	
20	The Role of Information Technology in Patient Engagement	155
	Sima Marzban, Paul Meade, Marziye Najafi, and Hossein Zare	
21	Ambient Intelligence Technologies for Visually Impaired: A Mapping Study	163
	Juliana Damasio Oliveira, João A. L. de Moraes Junior, and Rafael H. Bordini	
22	Voice for the Voiceless: Developing a Low-Cost Open-Source Communication Device for the Speech Impaired	169
	Travis Smith and Vasilios Pappademetriou	
23	Integration of Bioinformatics and Clinical Data to Personalized Precision Medicine	179
	Flavielle Blanco Marques, Gabriel Fernandes Leal, Giovanni Nicolas Bettoni, and Osmar Norberto de Souza	
24	Modeling the COVID-19 Epidemic in a Parallelized City Simulation	185
	Derek Stratton, William Garner, Terra Williams, and Frederick C. Harris Jr.	
Part V Management and Applications		
25	Techniques and Tools for Selection and Strategic Alignment of Projects and Projects Portfolio Balancing: A Systematic Mapping	195
	Djenane C. S. dos Santos, Adler D. de Souza, and Flávio B. S. Mota	

26	Methods for Detecting Fraud in Civil and Military Service Examinations: A Systematic Mapping	203
	Roberto Paulo Moreira Nunes, Rodrigo Bonacin, and Ferruccio de Franco Rosa	
27	Citizens Engagement in Smart Cities: A Systematic Mapping Review	209
	Rafael Leite, Adler Diniz, and Melise De Paula	
28	Use of Crowdsourcing Questionnaires to Validate the Requirements of an Application for Pet Management	215
	Vitor S. Vidal, Marco Aurélio M. Suriani, Rodrigo A. S. Braga, Ana Carolina O. Santos, Otávio S. Silva, and Roger J. Campos	
29	Discovery of Real World Context Event Patterns for Smartphone Devices Using Conditional Random Fields	221
	Shraddha Piparia, Md Khorrom Khan, and Renée Bryce	
30	Computation at the Edge with WebAssembly	229
	Jebreel Alamari and C. Edward Chow	
31	Analysis of Traffic Based on Signals Using Different Feature Inputs	239
	Baby Shalini Ravilla, Wolfgang Bein, and Yazmin Elizabet Martinez	
Part VI Theory and Computation		
32	Effect of Boundary Approximation on Visibility	247
	Laxmi Gewali and Samridhi Jha	
33	A Method for Improving Memory Efficiency of the Reachability Graph Generation Process in General Petri Nets	255
	Kohei Fujimori and Katsumi Wasaki	
34	An Evaluation for Online Power Down Systems Using Piece-Wise Linear Strategies	265
	James Andro-Vasko and Wolfgang Bein	
35	Mobile Test Suite Generation via Combinatorial Sequences	273
	Ryan Michaels, Md Khorrom Khan, and Renée Bryce	
36	Usability Smells: A Systematic Review	281
	Jonathan de Oliveira T. Souza, Adler Diniz de Souza, Leandro G. Vasconcelos, and Laercio A. Baldochi	
Part VII High Performance Computing Architectures		
37	Motivating Computer Science Students in Lower-Division Undergraduate Curriculum	291
	Yun Tian, Saqer Alhloul, Fangyang Shen, and Yanqing Ji	
38	Evaluation of Power Consumption and Application Optimization for Adaptive-Ticks Feature in Linux Kernel	297
	Abdullah Aljuhni, Shaji Yusuf, C. Edward Chow, Oluwatobi Akanbi, and Amer Aljaedi	
39	Sequence Alignment Algorithms in Hardware Implementation: A Systematic Mapping of the Literature	307
	Lucas S. M. Bragança, Adler D. Souza, Rodrigo A. S. Braga, Marco Aurélio M. Suriani, and Rodrigo M. C. Dias	

40	Hardware Logic Library and High-Level Logic Synthesizer Combining LOTOS and a Functional Programming Language	313
	Katsumi Wasaki	
41	GPU Acceleration of Sparse Neural Networks	323
	Aavaas Gajurel, Sushil J. Louis, Rui Wu, Lee Barford, and Frederick C. Harris Jr.	
42	Parallelizing the Slant Stack Transform with CUDA	331
	Dustin Barnes, Andrew McIntyre, Sui Cheung, John Louie, Emily Hand, and Frederick C. Harris Jr.	

Part VIII Social Computing/E-Learning

43	Recommender Systems Evaluator: A Framework for Evaluating the Performance of Recommender Systems	339
	Paulo V. G. dos Santos, Bruno Tardiolo Kuehne, Bruno G. Batista, Dionisio M. Leite, Maycon L. M. Peixoto, Edmilson Marmo Moreira, and Stephan Reiff-Marganiec	
44	Visualization of Georeferenced Data Through the Web: A Systematic Literature Review	347
	Lucas Lamounier Gonçalves Duarte and Adler Diniz de Souza	
45	Cognitive Issues in Intelligent Modeling of Pedagogical Task	355
	Marina Lapenok, Anna Lozinskaya, and Vasilisa Likhacheva	
46	Immersive Virtual Reality and Its Use in Developing Empathy in Undergraduate Students	361
	Éder Estrada Villalba and Fausto Abraham Jacques-García	
47	E-NEST Remote Learning Transition in STEM Education	367
	Fangyang Shen, Janine Roccasalvo, Jun Zhang, Yun Tian, Yang Yi, Yanqing Ji, Ashwin Satyanarayana, Xiangdong Li, Ahmet Mete Kok, Annie Han, and Hon Jie Teo	
48	Ethics and Human Values in the Software Design	373
	Alejandra Acuña, César Collazos, and Cristian Barría	

Part IX Pandemic

49	Using UAV, IoMT and AI for Monitoring and Supplying of COVID-19 Patients	383
	A. J. Dantas, L. D. Jesus, A. C. B. Ramos, P. Hokama, F. Mora-Camino, R. Katarya, O. P. Verma, P. K. Gupta, G. Singh, and K. Ouahada	
50	A Comprehensive Analysis of SARS-CoV-2 in India	387
	Debdeep Dey, Sarangi Patel, Karasani Tharun Kumar Reddy, and Siddharth Sen	
51	Virtual Hospital: A System for Remote Monitoring of Patients with COVID-19	397
	Vanessa Stangherlin Machado Paixão-Cortes, Walter Ritzel Paixão-Cortes, Dorval Thomaz, Felipe de Siqueira Zanella, Ricardo Luís Ravazzolo, and Gerson Luis da Silva Laureano	
52	Single-Cell RNA Sequencing Data Imputation Using Deep Neural Network ..	403
	Duc Tran, Frederick C. Harris Jr., Bang Tran, Nam Sy Vo, Hung Nguyen, and Tin Nguyen	

Part X Blockchain Technology

53 Blockchain and IoT: A Systematic Literature Review for Access Control Issues	413
André Mury de Carvalho, Bruno Guazzelli Batista, and Adler Diniz de Souza	
54 A Bitcoin Wallet Security System (BWSS)	421
Ibrahim Alkhamash and Waleed Halboob	
55 Disruptive Technologies for Disruptive Innovations: Challenges and Opportunities	427
Amjad Gawanmeh and Jamal N. Al-Karaki	
56 Framework for Securing Automatic Meter Reading Using Blockchain Technology	435
Esraa Dbabseh and Radwan Tahboub	

Part XI Biometrics, Pattern Recognition and Classification

57 Using Machine Learning to Process Filters and Mimic Instant Camera Effect	445
Deirdre Chong, John Farhad Hanifzai, Hassan Adam, Jorge Garcia, and Jorge Ramón Fonseca Cacho	
58 Benchmarking Accuracy and Precision of the Convolutional Neural Networks for Face Recognition on Makeup and Occluded Images	451
Stanislav Selitskiy, Nikolaos Christou, and Natalya Selitskaya	
59 Combined Classification Models Applied to People Personality Identification	457
Flávio Mota, Melise Paula, and Isabela Drummond	
60 A Health Detection Model Based on Facial Data	463
Sunil Manzoor and Shahram Latifi	
61 Performance Comparison of Algorithms Involving Automatic Learned Features and Hand-Crafted Features in Computer Vision	469
Rocky Y. Gonzalez and Shahram Latifi	

Part XII Data Sciences

62 Big Data Analytics in Social Media: A Triple T (Types, Techniques, and Taxonomy) Study	479
Md. Saifur Rahman and Hassan Reza	
63 CARS: A Containerized Amazon Recommender System	489
Adam Cassell, Andrew Muñoz, Brianna Blain-Castelli, Nikkolos Irwin, Feng Yan, Sergiu M. Dascalu, and Frederick C. Harris Jr.	
64 Using Technologies to Uncover Patterns in Human Trafficking	497
Annamaria Szakonyi, Harshini Chellasamy, Andreas Vassilakos, and Maurice Dawson	
65 Data-Driven Identification of Pedagogical and Curricular Factors Conducive to Student Satisfaction	503
Laura Sorto, Sourav Mukherjee, and Vasudevan Janarthanan	

66	An Information Quality Framework for College and University Websites	509
	Joseph Elliot and Daniel Berleant	
67	An Agile MDD Method for Web Applications with Modeling Language	519
	Breno Lisi Romano and Adilson Marques da Cunha	
Index	527

Chair's Message



Welcome to the 18th International Conference on Information Technology – New Generations – ITNG 2021. This year, due to the pandemic, we have witnessed a big decline in attendance for international conferences and meetings. COVID-19 has affected our lives in many ways, resulting in economic downturn, budget cuts, travel restrictions, and confinements. On the other hand, on a positive note, the pandemic has caused a boost in R&D in all the technologies that can make remote operation possible.

It is a pleasure to report that despite the virus-related problems, we have another successful year for our conference. Our conference attracted many quality submissions globally. The papers were reviewed for their technical soundness, originality, clarity, and relevance to the conference. The conference enjoyed expert opinion of over 40 authors and non-author scientists who participated in the review process. Each paper was reviewed by at least two independent reviewers. At the end, 67 papers were accepted to shape the ITNG 2021 program.

The chapters in this book address the most recent advances in such areas as Big Data Analytics, Cybersecurity, Data Mining, e-Health, High Performance Computing, IoT & CPS, Software Engineering, Social computing, and Biometrics. In addition to technical presentations by the authors, the conference features two keynote speakers.

Many people contributed to the success of this year's conference by organizing symposia or technical tracks for the ITNG. Dr. Doina Bein served in the capacity of conference vice chair. We benefited from the professional and timely services of major track organizers and associate editors, namely Drs. Azita Bahrami, Cristian Barra Huidobro, Doina Bein, Luiz Alberto Vieira Dias, Ray Hashemi, Kashif Saleem, Fangyan Shen, David Cordero Vidal, and Hossein Zare.

Others who were responsible for solicitation, review, and handling the papers submitted to their respective tracks/sessions include Drs. Wolfgang Bein, Poonam Dharam, and Mei Yang.

The help and support by Springer in preparing the ITNG proceedings is specially appreciated. Many thanks are due to Michael Luby, senior editor and supervisor of publications, and Brian Halm, production editor at Springer, for the timely handling of our publication order. We also appreciate the efforts made by Springer Project Coordinator Olivia Ramya Chitranjan. Olivia spent much time looking very closely at revised articles to make sure they are formatted correctly according to the publisher's guidelines.

We also thank the technical assistance of Prabhas Kumra in setting up the conference program and Zoom communication for us. Finally, the great efforts of the conference secretary, Ms. Mary Roberts, who dealt with the day to day conference affairs, including timely handling volumes of emails, are acknowledged.

I hope that you all enjoy the ITNG 2021 program and find it technically and socially fulfilling.

Shahram Latifi
The ITNG General Chair

ITNG 2021 Reviewers

Acuña, Alejandra	Janine, Roccosalvo
Alatab, Ahmed	Khan, Farhan
Alwady, Ali	Latifi, Shahram
Andro-Vasko, James	Mahto, Rakeshkumar
Bahrami, Azita	Marques, Johnny
Barría, Cristian	Mialaret, Lineu
Bein, Doina	Montini, Denis
Bein, Wolfgang	Ortega, Saul
Chu, Elaine	Saleem, Kashif
Cordero, David	Shen, Fangyang
Derhab, Abedulahed	Soto, Daniel
Dharam, Poonam	Vieira-Dias, Luiz-Alberto
Galeazzi, Lorena	Yang, Mei
Gawanmeh, Amjad	Yenny, Mendez
Ghaleb, Mukhtar	Zare, Hossein
Gofman, Mikhail	Zhong, Danping
Hashemi, Ray	

Part I

AI and Robotics

Conceptualisation of Breast Cancer Domain Using Ontology

Reshmy Krishnan, P. C. Sherimon, and Menila James

Abstract

The conceptualization of the breast cancer domain using ontology is an emerging area of intelligent decision support system. Eventhough this is not the replacement for clinicians, this intelligent system can support them in an effective way during diagnosis. As the system requires data from clinicians and patients, the unorganized data is gathered and processed. As the input data are unstructured, it is hard to gather information and share knowledge from that. An adaptive questionnaire is used to gather data to optimize the result of the system. The paper discusses the prototype model which uses various ontology as the knowledge base, java engine to provide information to modeller and a reasoner to take effective decision. SPARQL is used to retrieve required information as per the conditions. Protégé that supports OWL representation provides a platform to build concepts and relationships. How ontology is representing the details of Breast cancer guidelines and how instances of the class are identified using the query are shown in this paper.

Keywords

Breast cancer · Ontology · Description logic (DL) · Protégé · Semantic medical profile · Questionnaire · SPARQL · OWL · Reasoner · Decision support system

R. Krishnan (✉) · M. James
Department of Computing, Muscat College, Muscat, Sultanate of Oman

e-mail: reshmy@muscatcollege.edu.om;
menila@muscatcollege.edu.om

P. C. Sherimon
Department of IT, Arab Open University, Muscat, Sultanate of Oman
e-mail: sherimon@aou.edu.om

1.1 Introduction

When the growth of cells are out of control in the breast cells, cancer is initiated [1].

1.1.1 The Intelligent Decision Support System and Its Role in Clinical Domain

The system of Analytical techniques along with Interaction with user is used to develop decisions for semi structures decision problems [3]. The integration of artificial Intelligence (AI) with these systems enhances and assist decision maker and is called Intelligent decision support systems (IDSS). “Intelligent decision support is provided by a system that helps in decision-making through a display of intelligent behavior that may include learning and reasoning” [7]. Such learning and reasoning can be achieved through implementing rule-based expert systems, knowledge-based systems or neural network systems (Fig. 1.1).

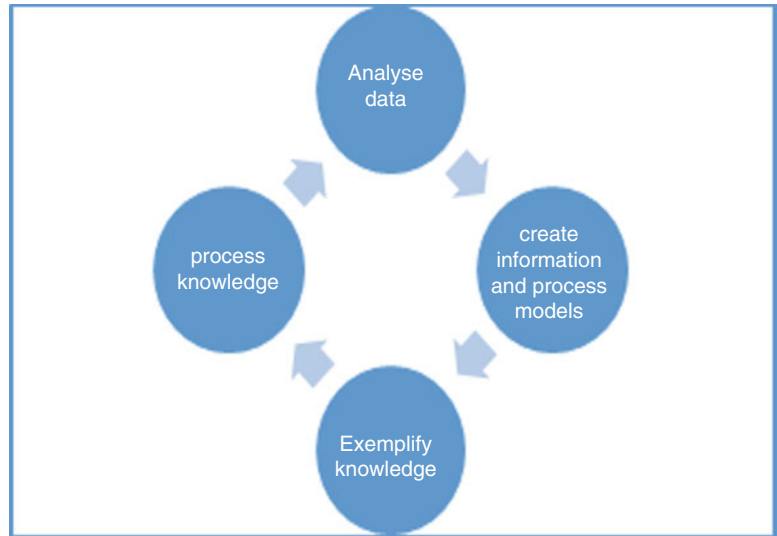
IDSS can be used in clinical decision support systems in many ways such as

- Diagnose the patient data through regular monitoring and interpretation
- Support disease management through alerts
- Help public health by predicting epidemic diseases.
- Support clinicians to decide an accurate diagnosis.

1.1.2 Knowledge Base and Ontology

Knowledge base (KB) is the collection of information which analyses many relations and inferring relations between them unlike to database which builds structure to collect data [4]. As ontologies can define the link between various types of

Fig. 1.1 The knowledge cycle with AI methods and tools



semantic knowledge, they can be used in data searching strategies [5].

Ontology is a conceptualization of shared domain. The object oriented concept is used in the development of ontologies where classes and objects are specified under a domain [1]. Domain of ontology can cover a wide variety of topic such as medical, agriculture, education, health etc. with concise terminologies and pictorial representation that provides its application in research fields. The query terms and members of ontology are mapped and the result is produced [2]. Thus ontological approach towards medical domain by using consolidated knowledge base and patient profile can be utilized for risk prediction of various diseases [3].

Breast cancer being a deadly disease among females in most part of the world has given a major concern for the early detection to reduce the death rates. In spite of technologies and developed medical field, mortality rates due to breast cancers are uncontrolled even in developed country like Oman. Cancerous tumors are malignant which invade nearby tissues. To find whether the tumors are malignant or not, researches have done by using ontology based object oriented case-based reasoning frameworks [6]. According to researches, out of 122 patients with invasive breast cancer in Oman, 119 were females 54.9% of which were in the age group 41–60 followed by 32% in age group below 40 [7]. The Oman cancer Association screened 8278 women where predictors of compliances were family history of breast cancer and self-examination [8]. The proposed project focuses on the prediction of breast cancer by using clinical guidelines and gathered information about symptoms from patients of Oman.

The outline of this paper is designed like this: Introduction followed by literature survey, proposed system architecture. Preliminary results and discussion, results and conclusion and future works.

1.2 Literature Review

Relations and concepts of breast cancer clinical guidelines, demographics, test results, treatments and pathologies are curated by Breast Cancer Ontology in Clinical Decision.

Clinical Reasoning ontology (CRO) based CDSS identifies medical knowledge and reasoning concepts and extracts these concepts and properties [10].

In paper [11], breast masses are diagnosed using expert concepts and rules where it is classified into benign cancer or malignant. Machine learning tool is used for the semantic annotation.

Web Ontology Language (OWL) is a computational logic-based language designed to author ontologies [12]. Description Logic (DL) serves formal knowledge representation for the description of concepts and roles. Health record and knowledge source linkage is done by the structured representation of clinical records, dynamic coding and matching of clinical context with knowledge source.

The paper [13] describes the creation of ontologies as a shared domain conceptualization through ontology-driven conceptual modeling and proposes an enterprise that represents ontology instances by using the open source software Protégé. The technical feasibility and wide area application of the software is explained, and reasoning capability of the system is shown as an example.

1.3 Proposed System and Architecture

The proposed system is the Breast Cancer Prediction System based on ontology containing three ontologies namely questionnaire ontology, clinical guidelines ontology and symptom ontology. Figure 1.2 depicts the proposed architecture. The system follows the methods of an expert system which

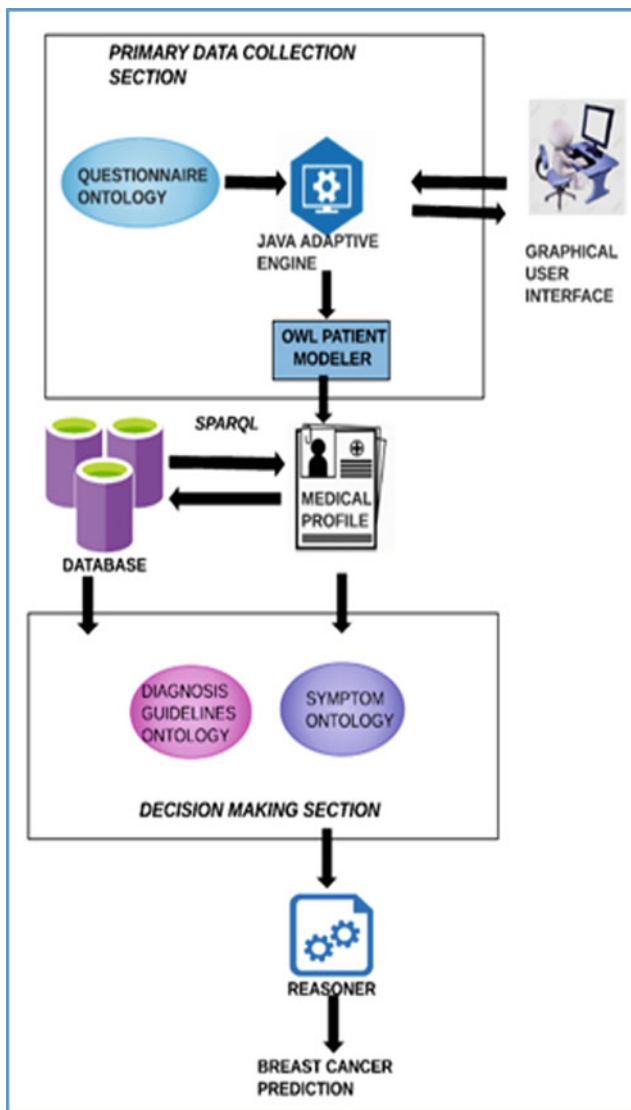


Fig. 1.2 Architecture of ontology based breast cancer prediction system

creates questionnaires to collect information of the patient such as demography- basic and medical, family history, symptoms, test results etc. and combines it with guidelines and symptom information to predict the risk and stage.

As discussed earlier, the ontology driven adaptive questionnaire in primary information collection section, diagnosis guidelines ontology and symptom ontology in decision making section forms the main components of the system.

The Graphical User Interface (GUI) provides an interface with questionnaires to the user through which relevant information is collected and results are displayed to users that are clinicians and patients. The interface is developed using java for more user friendly appearance.

Primary data collection module uses questionnaire ontology and OWL modeller by the Protégé software to auto-

matically generate patient's semantic medical profile. The patient's semantic medical profile is formulated by the use of adaptive questionnaire [14] which aims at collection of information with dynamic modification in response to the user [15, 16].

Interpreting descriptive annotations like symptoms, findings and observations provide base to clinical decision making with symptom ontology containing their information [17]. The symptom ontology contains symptoms and their relations that help the clinicians to identify and relate the chances and risk of breast cancer [19].

Risk factors includes proliferative abnormalities of the breast in lobular and ductal epithelium including hyperplasia, atypical hyperplasia and invasive carcinoma according to Clinical Practice Guidelines in Oncology by NCCN [18] provides the base for diagnosis guidelines ontology.

1.4 Preliminary Results and Discussion

1.4.1 Breast Cancer Ontology in Protégé 5.5

Protégé is a free, open-source software developed by Stanford University that provide a platform to build intelligent decision-making system through classes, relations and query formation. The software was aimed mainly towards the beginners for the easy creation of ontology.

The implementation process of the proposed systems involves five steps:

1.4.1.1 Determining the Domain of Ontology

Ontology can be applied to a vast number of domains according to the requirement. For the creation of ontology, determination of domain and fixation of relationships to store and extract data in the OWL-based knowledge base is the fundamental procedure. As the project deals with the early detection of breast cancer, Breast Cancer is served as the domain of our ontology.

1.4.1.2 Obtaining Class Hierarchy by Means of Tools in Protégé-OWL

The ontology concepts, relationship between these concepts and constraints of relationships are built. The graphical view of hierarchy of classes in the ontology is shown in Fig. 1.3. Visualising ontology provides an alternative way of navigating and exploring the models. Protégé provides some useful ontology visualisation tools namely, OntoGraf and OWLViz.

Unlike OWLViz, the neighbouring classes based on relationships can be viewed by selecting a class or subclass in visualisation panel of OntoGraf. Figure 1.4 shows the visualisation of class hierarchy using OntoGraf tool. The OWLViz visualisation is shown in Fig. 1.5.

The initial structure includes classes for patient as the first and foremost step is to gather the information about the patient's demography and family history. The symptom class consists of subclasses external and internal with instances of basic symptoms like nipple discharges, lumps, rashes, irregular menstruation. This provides a rough idea about the ontological approach towards breast cancer.

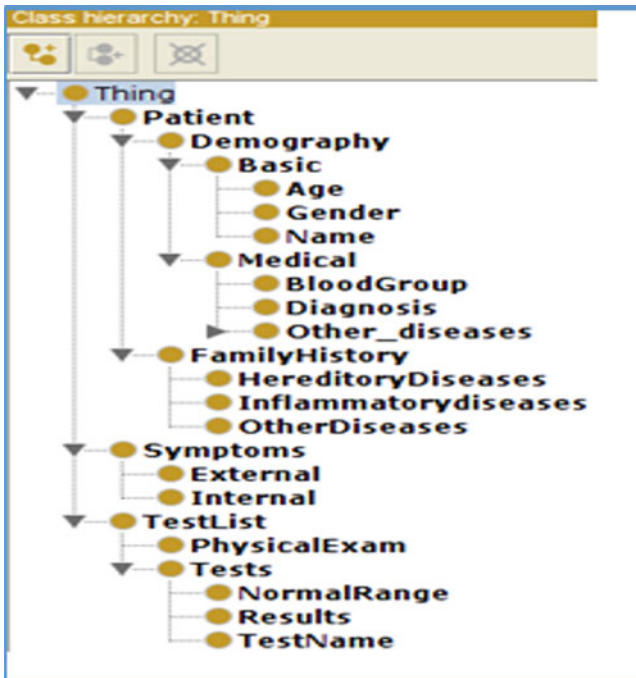
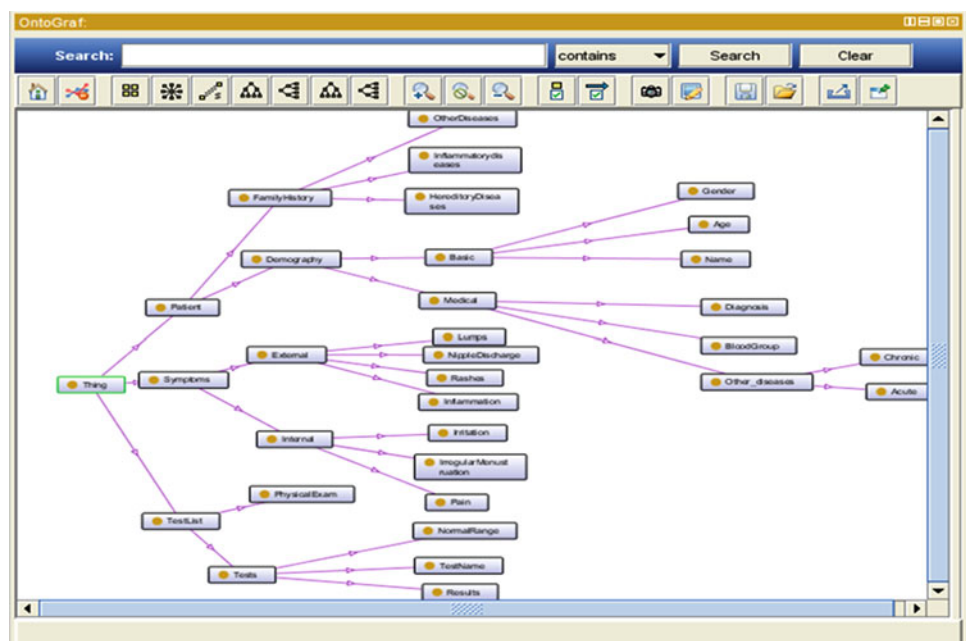


Fig. 1.3 Classes and subclasses of breast cancer ontology

Fig. 1.4 Class hierarchy in OntoGraf



1.4.1.3 Object and Data Property Creation

The object property is used to create relationship between different classes to connect them for querying. The created object properties of classes of Breast Cancer ontology is given as Fig. 1.6.

Another property that determines the link between individuals and data type is data property which is shown in Fig. 1.7. The domain of the data property is selected from the class hierarchy and range indicates the data type of the properties [20].

1.4.1.4 Querying Using SPARQL

Extraction of information from large datasets is made easy by a tool in Protégé called SPARQL query language [21]. SPARQL is similar to SQL but it has high performance Query to obtain the answer of the question “How many patients have the symptom lumps or other diseases?” is shown in Fig. 1.8.

After execution, the result will be the number of patient patients having symptom lumps or other diseases related to breast cancer.

1.4.1.5 Validating Ontology by Comparing Query Result with Ontology Instances

When the query results match the instances of the concepts in ontology, it can be identified that the ontology is validated [21].

1.4.2 Results

Query generation and searching interface in protege is provided by DLQuery Tab. The Fig. 1.9 shows a basic query to find instances of a class.

Fig. 1.5 Class hierarchy in OWLViz

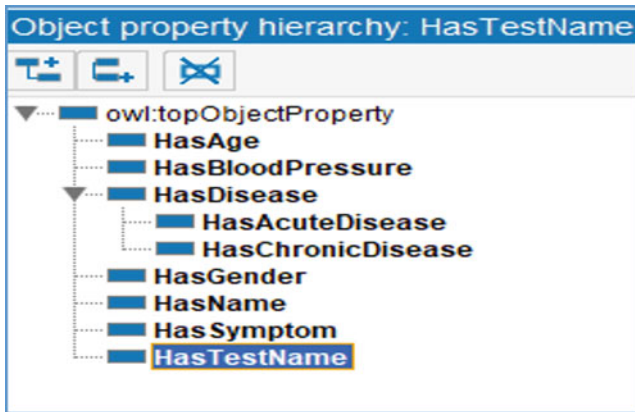
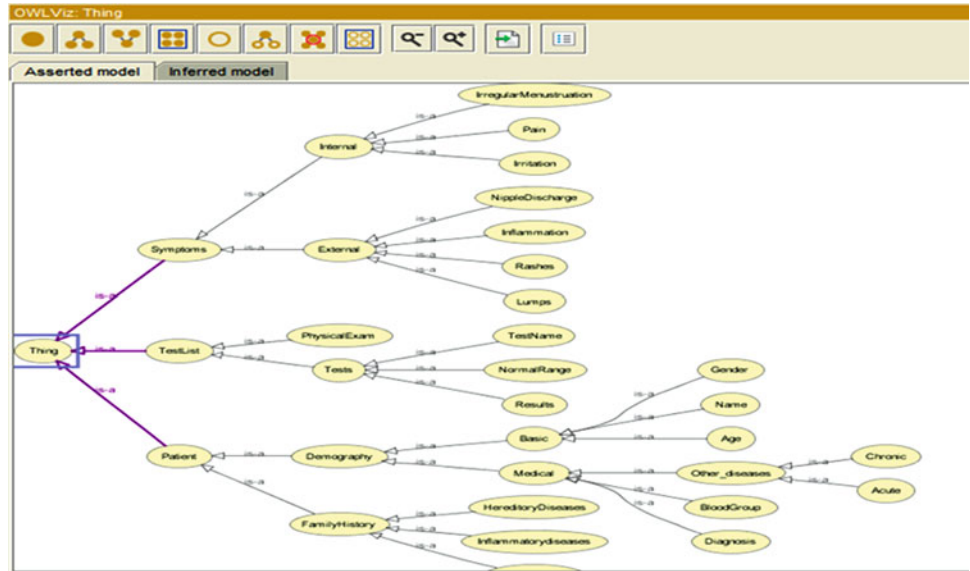


Fig. 1.6 Object property hierarchy

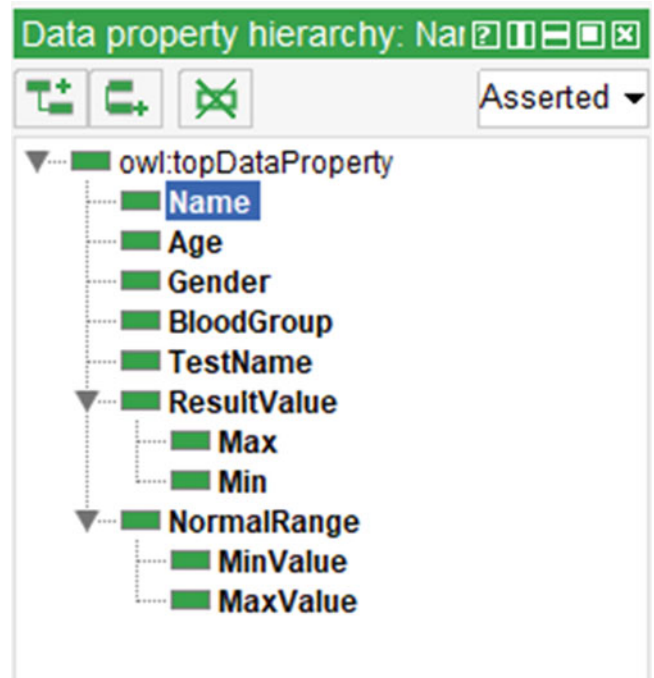


Fig. 1.7 Data property hierarchy

1.5 Conclusion and Future Work

The work is aimed to help the physicians in the prediction of breast cancer by using multidisciplinary knowledge gathered over the knowledge bases. With the wide usage of this system early detection of breast cancer can be made easier as knowledge sharing is formulated through the collection of direct information about symptoms, test results and family history from the patient.

Extending the standardized ontologies to represent the knowledge of breast Cancer domain is important for the improvement of accuracy. Ontology driven adaptive questionnaire, ontologies on guidelines and symptoms will be explained in the future works. The adaptive questionnaire will be given to patients and the prediction depends on the

defined symptom ontology and test strategies combined with the collected information.

Further updating and correction of defined ontologies is possible and it can be implemented with simple alterations according to any new invention or research. Thus the proposed system can be leveraged as it is lucrative for the extensive application by capturing the knowledge of the Breast Cancer domain.

Fig. 1.8 Querying using SPARQL

```
SPARQL query:
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
WHERE
{
  {?Patient ab:HasSymptom ?Lumps}
  UNION
  {?Patient ab:HasDisease?OtherDiseases}
}
```

Fig. 1.9 Basic query in DL Query

The screenshot shows a web-based interface for a DL query. At the top, there is a text input field containing the word "External". Below the input field are two buttons: "Execute" and "Add to ontology". The main area is titled "Query results" and is divided into several sections:

- Equivalent classes (1):** A single entry "External" with a question mark icon.
- Ancestor classes (2):** Two entries: "Symptoms" and "Thing", each with a question mark icon.
- Super classes (1):** A single entry "Symptoms" with a question mark icon.
- Instances (4):** Four entries: "Lumps", "Rashes", "IrregularMenstruation", and "NippleDischarge", each with a question mark icon.

On the right side of the "Query results" section, there is a vertical list of checkboxes for filtering the results:

- Super classes
- Ancestor classes
- Equivalent classes
- Subclasses
- Descendant classes
- Individuals

Acknowledgement This paper is part of funded Project “An Intelligent Clinical Decision Support System for Breast Cancer in Sultanate of Oman” from Research Council, Sultanate of Oman in call TRC/BF-P/MC/01/2018.

References

1. S. Viadomonte, F. Burstein, Chapter 4: From knowledge discovery to computational intelligence: A framework for intelligent decision support systems, in *Intelligent Decision-Making Support Systems*, (Springer-Verlag London Limited, London, 2006), pp. 57–78
2. R. Basu, U. Fevrier-Thomas, K. Sartipi, Incorporating hybrid CDSS in primary care practice management. McMaster eBusiness Research Centre, November 2011
3. V.L. Patel, E.H. Shortliffe, M. Stefanelli, P. Szolovits, M.R. Berthold, R. Bellazzi, A. Abu-Hanna, The coming of age of artificial intelligence in medicine. *Artif. Intell. Med.* **46**(1), 5–17 (2009)
4. M. Alfonso, M.M. Aref, A.-B.M. Salem, An ontology-based system for cancer diseases knowledge management. *Int. J. Inf. Eng. Electron. Bus.* **6**(6), 55–63 (2014)
5. D. Parry, A fuzzy ontology for medical document retrieval, in *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation-Volume 32*, (Australian Computer Society, Inc., 2004)
6. P.C. Sherimon et al., Ontology based system architecture to predict the risk of hypertension in related diseases. *J. Inf. Process. Manag.* **4**(4), 44–50 (2013)
7. A. Bish et al., Understanding why women delay in seeking help for breast cancer symptoms. *J. Psychosom. Res.* **58**(4), 321–326 (2005)
8. M. Sewak et al., SVM approach to breast cancer classification, in *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, (IEEE), p. 2007
9. E.A.M.L. Abdrabou, A.E.-B.M. Salem, A breast cancer classifier based on a combination of case-based reasoning and ontology approach, in *Proceedings of the International Multiconference on Computer Science and Information Technology*, (IEEE, 2010)
10. S. Kumar et al., Changing trends of breast cancer survival in sultanate of Oman. *J. Oncol.* **2011**, 316243 (2011)
11. S. Al Balushi, Predictors of compliance and predictive values of the breast cancer screening program of the Oman Cancer Association (2009–2016). (2017)

12. M. Gong et al., Toward early diagnosis decision support for breast cancer: Ontology-based semantic interoperability. *J. Clin. Oncol.* **37**, e18072 (2019)
13. P.I. Dissanayake, T.K. Colicchio, J.J. Cimino, Using clinical reasoning ontologies to make smarter clinical decision support systems: A systematic review and data synthesis. *J. Am. Med. Inform. Assoc.* **27**(1), 159–174 (2020)
14. S.T.B. Ameer et al., Ontology based decision system for breast cancer diagnosis, in *Tenth International Conference on Machine Vision (ICMV 2017)*, vol. 10696, (International Society for Optics and Photonics, 2018)
15. J.J. Chelsom, N. Dogar, Linking health records with knowledge sources using OWL and RDF. *ITCH*. (2019)
16. B. Reitemeyer, H.-G. Fill, Ontology-driven enterprise modeling: A plugin for the Protégé platform, in *Enterprise, Business-Process and Information Systems Modeling*, (Springer, Cham, 2019), pp. 212–226
17. M.-M. Bouamrane, A. Rector, M. Hurrell, Gathering precise patient medical history with an ontology-driven adaptive questionnaire, in *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, (IEEE, 2008)
18. P.C. Sherimon, P.V. Vinu, Y. Takroni, R. Krishnan, Developing a survey questionnaire ontology for the decision support system in the domain of hypertension. *IEEE South East Conference*, April 2013
19. P.C. Sherimon et al., Adaptive questionnaire ontology in gathering patient medical history in diabetes domain, in *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, (Springer, Singapore, 2014)
20. H. Oberkampff et al., Interpreting patient data using medical background knowledge. *ICBO***897**(3), 1–5 (2012)
21. R.W. Carlson et al., Breast cancer: Noninvasive and special situations. *J. Natl. Compr. Canc. Netw.* **8**(10), 1182–1207 (2010)

Traffic Light Control and Machine Learning: A Systematic Mapping Review

2

Dimitrius F. Borges, Edmilson M. Moreira, Adler D. de Souza, and João Paulo R. Leite

Abstract

The global vehicle fleet has grown rapidly over the past decade, impacting the way traffic is to be managed. Vehicle traffic management and control through technology is a well-known and widely studied problem that continues to present challenges and opportunities for action, mainly due to the growing demand, the mentioned increase in the vehicle fleet, and inefficiency of current systems, generally based on fixed-time traffic lights. Solutions have been presented for this scenario, and among them, Artificial Intelligence (AI) and Machine Learning (ML) techniques have stood out. The AI/ML field, however, is vast and varied. This article proposes a survey of the most used AI/ML techniques in the management of vehicular traffic lights, and it does so through a Systematic Mapping Review (SMR), pointing out models that receive greater focus, research trends and gaps.

Keywords

Artificial intelligence · Machine learning · Systematic review · Traffic control · Reinforcement learning · Vehicles · Decision-making · Trends · Gaps

2.1 Introduction

Brazilian vehicle fleet grew 21.5% between 2012 and 2017 [6], and similar situations can be observed in Europe [1]. This growth leads to a constriction of traffic lights, turning street intersections into bottlenecks, reducing traffic flow

D. F. Borges · E. M. Moreira (✉) · A. D. de Souza · J. P. R. Leite
Universidade Federal de Itajubá, Itajubá, Brazil
e-mail: edmarmo@unifei.edu.br; adlerdiniz@unifei.edu.br;
joaopaulo@unifei.edu.br

and increasing the average waiting time in line, and, hence, the dissatisfaction of drivers. This fact is enhanced by a characteristic of the most common type of traffic lights in use today: fixed operating times, which do not take into account the variation in demand.

Vehicular traffic systems can be modeled, with data collected, mapped, and transferred to a computer system. Artificial Intelligence (AI) techniques can then be applied to them, supporting the decision-making process and achieving better results such as reduced vehicle queue size (*i.e.*, lower waiting time) and increased vehicle flow. These decisions are mainly based on the control of traffic light times, which are then adapted to the context and the sensed environment.

Among the AI techniques that can be used for this purpose, Machine Learning (ML) stands out, which aims to provide computer systems with the ability to acquire knowledge inferred from raw data [8]. This type of learning, characterized by induction (*i.e.*, learning by example), allows computers to solve problems that involve real-world knowledge with a minimum of human interference.

Given the premise of ML and the characteristics of a traffic control system, it is expected that there will be studies that employ the former to solve the latter. Besides, there are several models and lines of research in ML, and many of them can be applied in this context. Therefore, building a literature review through systematic mapping emerges as a useful, if not necessary, endeavor for researchers in the field. This article proposes carrying out this work, structuring an overview of the main AI and ML models and techniques used to solve the problems and dissatisfactions resulting from the growing vehicular flow in our cities.

The paper is organized as follows: Sect. 2.2 provides a brief context on the problem, and Sect. 2.3 describes the methodology, including research questions and search expressions. In Sects. 2.4 and 2.5, we present and discuss the results, and Sect. 2.6 contains our conclusions.

2.2 Context

Artificial Intelligence (AI) is a branch of computing that seeks to emulate human intelligence through computational systems [5]. These techniques have been proposed as solutions for vehicle traffic control for decades [4], as intelligent agents that, with power to capture sensitive data in the environment and make decisions, are able to determine the best course of action [15].

Taking into account the large number of AI/ML models, a systematic mapping study becomes useful as it allows the visualization of a complete picture of what has been most used to solve the characteristic problems of this scenario, which techniques stand out and which have been neglected or underestimated.

2.3 Methodology

A systematic mapping review (SMR) provides a structure on the type and results of published research on a topic, categorized and organized in a concise representation of these results [12]. Our objective is to analyze experience reports and scientific publications through a systematic mapping, to identify resolution proposals and study trends for AI applications in vehicular traffic control, concerning ML models capable of predicting and manage vehicular flow, in the academic and industrial context.

In this SMR, the main research question is:

MQ: *What are the main ML models used in vehicular traffic control?*

The main question was supplemented by the following secondary questions, concerning each model:

SQ1: *Does it have a learning structure?*

SQ2: *Is it applied to isolated intersections or a network of intersections?*

SQ3: *Is it applied to a real case?*

A conclusive question was also proposed:

CQ: *What are the trends and gaps for this field of study?*

2.3.1 Restrictions, Inclusion and Exclusion Criteria

The search was restricted to publications obtained from the selected databases and references of these publications. Publications must be in Portuguese or English and made available between 2006 and 2019. The selection of studies goes through two filters of inclusion or exclusion criteria, respecting the following nomenclature format: [C] criterion of [E or

Table 2.1 Inclusion and exclusion criteria for publications

Code	I or E	Description
CE1-01	E	Absence of selected keywords
CE1-02	E	Publications of workshops, keynote speeches, surveys, tutorials, courses and similar
CE1-03	E	Does not deal with vehicular traffic control
CE1-04	E	Duplicate articles
CE1-02	E	Publications that do not simulate or test the proposed model
CI1-01	I	Mention of ML model applied to vehicle traffic control
CI1-02	I	Mention of AI model applied to vehicular traffic control
CE2-CT	E	Does not propose AI or ML model to control vehicular traffic at intersections
CE2-TR	E	Model that does not apply changes in real time
CE2-CD	E	Model that does not collect data in real time
CE2-MO	E	Publications that do not specify the model used

I] Exclusion or Inclusion of the [1st or 2nd] filter—code XX. Table 2.1 shows the criteria.

2.3.2 Search Engines and Search Expressions

Scopus,¹ IEEE Xplore,² and ACM Digital Library³ were chosen as search engines and electronic databases. The main search phrase was:

(traffic signal control OR traffic signal controlling OR traffic signal management) AND (Machine Learning OR Deep Learning OR Artificial Neural Networks OR Artificial Intelligence) AND (model OR application)

A total of five search rounds were performed on the selected engines using the selected search phrase, ensuring results consistent with the intention of the search. In each of the rounds after the first, comparisons were made with a created control group, which gathers five articles considered essential, and that must be found in the other rounds.

2.4 Results

In the process, 327 articles were found and evaluated: 298 from search engines and 29 from *snowballing*, which consists of evaluating references from relevant articles in search of publications not previously identified. Two articles were selected for the process, [21] and [23]. Both were chosen because they are the work of *surveys*.

¹<https://www.scopus.com/>.

²<https://ieeexplore.ieee.org>.

³<https://dl.acm.org/>.

The preliminary selection took place when analyzing “Titles” and “Abstracts” of each of the articles, according to the research questions. In a second step, these articles were reevaluated by the inclusion and exclusion criteria. If “Titles” and “Abstract” were insufficient for decision making, “Introduction” and “Conclusion” sections were also analyzed.

After the entire process, depicted in Fig. 2.1, 132 articles were selected for the study. Data extracted from the selected publications and stored in a database were: title, author(s), date of publication, how it was found, keywords, and observations.

Once the articles were selected, a classification process was carried out to create a brief description of each article. The process involved two steps: Identification of descriptive keywords and grouping of keywords, creating classes. The process culminated in the definition of the following classes: Artificial Neural Networks (ANN), Reinforcement Learning (RL), Meta-heuristics, Fuzzy Logic, Dynamic Programming (DP), Game Theory, Others (do not apply to previous classes), and, regarding their structure, systems were classified as Isolated (single intersection) or Networked (more than one agent assisting the decision process). The final classification can be seen in Table 2.2.

2.4.1 Analysis of Results

Figure 2.2 shows the proportions of the different classes. The dominance of Reinforcement Learning (RL) is clear

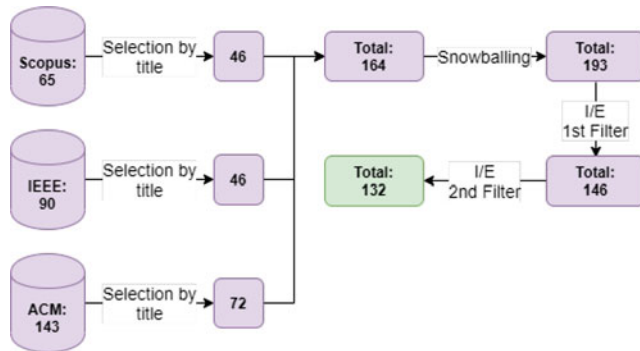


Fig. 2.1 Search process flow

Table 2.2 Classification of selected articles

Class	Isolated	Networked
Reinforcement learning	25	25
Neural networks	8	2
Meta-heuristics	19	6
Fuzzy	21	3
Dynamic programming	3	5
Game theory	4	2
Others	6	3

(37.9%), with values indicating that it is referenced twice as often as the second technique, Meta-heuristics (18,9%). Figure 2.3 corroborates this statement by showing that, over the years, RL has always been relevant, having at least one publication per year and, still, representing at least half of the publications in years 2011, 2012, 2013, 2014, 2016, and 2018.

Also, RL is the only one to show a relevant increase, indicating that the technique appears as a trend. It is worth noting the decline of fuzzy models, an indication that the technique has been losing its relevance in the field, giving way to others that are more promising.

Regarding the structure, as shown in Fig. 2.4, isolated systems are predominant. However, this trend does not apply to RL models, which present a perfect balance, different from what happens in other models that, except for DP, have a strong tendency to isolated architectures. The contrast probably lies in the intrinsic limitations of each type of model: Meta-heuristics require high processing time (maybe prohibitive), given the complexity of the system; fuzzy has a simpler premise, which makes it challenging to consider decisions made by other agents. RL, on the other hand, can be prepared, from its conception and training, to consider decisions made by other agents, being then relatively easier to apply it to a synchronous and functional networked architecture. It is impossible, though, to determine which architecture is most likely to dominate the future. A small downward trend was identified in the isolated models, as shown in Fig. 2.5, which does not reflect directly in an increase in the networked models. It may be a sign of a slight fluctuation in interest in this field and not an exchange between architectures.

With the predominance of RL models, a more in-depth analysis of the technique was carried out, seeking to identify the different approaches used. Five were found to be relevant: Q-learning [9], Deep Q-learning [8], SARSA [19], W-learning [7], and Actor Critic [14]. Articles were classified as shown in Table 2.3.

Figure 2.6 shows that Q-learning stands out (66%), especially if we consider that the second most used approach is Deep Q-learning (12%), a variation of the first. The reason for this lies in the fact that Q-learning is one of the most widely used methods in literature [10, 16] and has shown good performance in various fields. Even so, three other proposals were found, which might be an attempt to change focus, leaving dominant Q-learning to explore alternative and little known or even more complex proposals, which would benefit the scientific community. We cannot state, though, that such techniques try to fill some existing inefficiency, considering that this study did not raise them.

2.4.2 Answers to Research Questions

MQ: What Are the Main ML Models Used in Vehicular Traffic Control? The study shows the dominance of RL, with 37.9% of the publications evaluated. It was found that the most used RL technique is Q-learning, with 33 publications (66%),

more than any other technique or model, allowing us to affirm that RL with Q-learning is the most used ML model within the field.

SQL: Does It Have a Learning Structure? Despite the fact that our original proposal was to search for ML models,

Fig. 2.2 Proportion of publications by technique

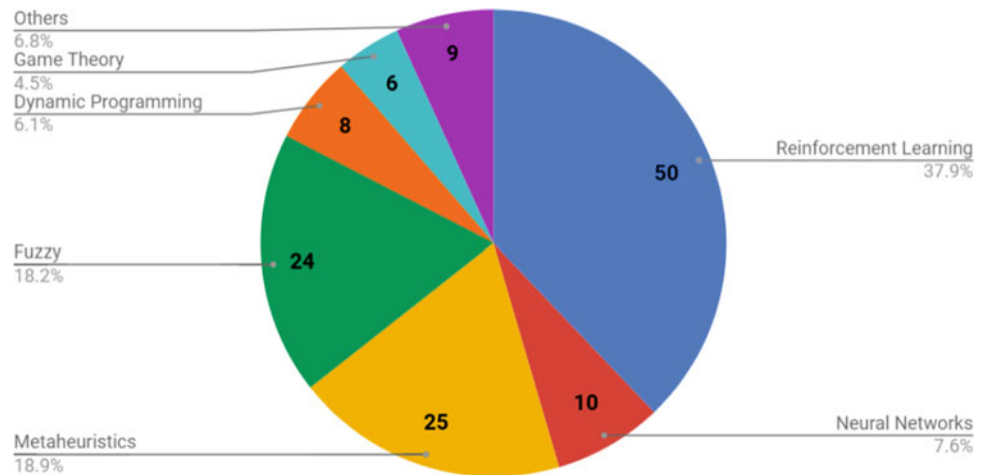


Fig. 2.3 Number of publications for each technique per year

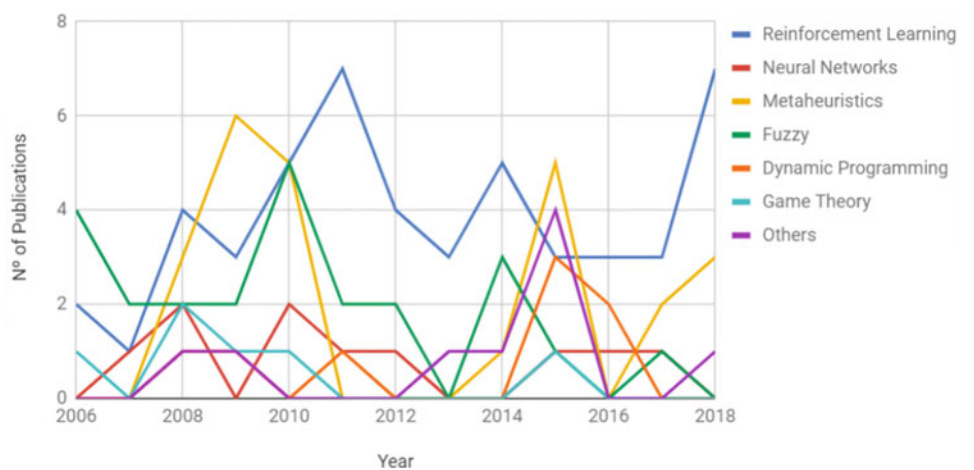


Fig. 2.4 Proportion between architectures by technique

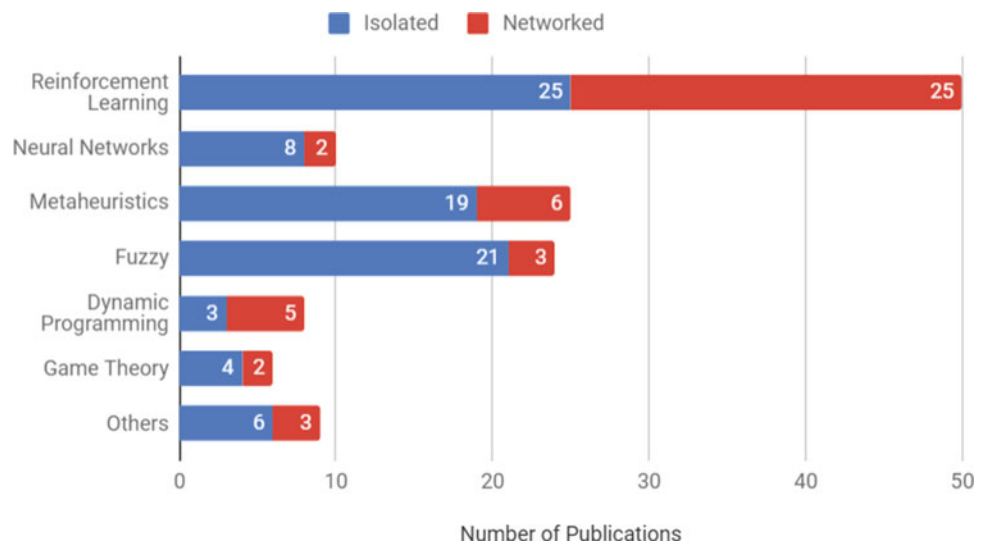


Table 2.3 Classification of reinforcement learning articles

Class	Isolated	Networked
Q-learning	14	19
Deep Q-learning	5	1
SARSA	4	-
W-learning	-	1
Actor critic	2	1
Others	-	3

results that did not directly involve this field of study were not discarded. Thus, Fuzzy Logic was the only AI technique found that does not have an intrinsic learning process - it can still be applied to the studied context, especially when combined with other models [13]. As shown in Fig. 2.3, though, fuzzy-based models have been falling into disuse.

SQ2: Is It Applied to Isolated Intersections or a Network of Intersections? Models were found that address the prob-

lem considering both networked systems and isolated ones. As expected, isolated architectures were shown to be more common than networked ones, given that the latter is naturally more complex. Furthermore, it was found that networked models have not gained significant popularity over the years (Fig. 2.5). Even when isolated models usage declined, there was no transfer of interest to networked systems, indicating that the decrease in popularity was possibly due to a mild loss of interest in the area itself and not in the system architecture. However, it is expected that with the spread of more powerful portable computer systems and AI frameworks, there might be a natural migration from the simpler model to the more complex (and complete).

SQ3: Is It Applied to a Real Case? All models found during our SMR were limited to simulation, with a single exception [18]. Our study does not offer enough information to point out the real reason behind this finding. The only indication is given by relating the complexity of the proposed models and the computational power available, that is, there is a distance

Fig. 2.5 Architectures over the years

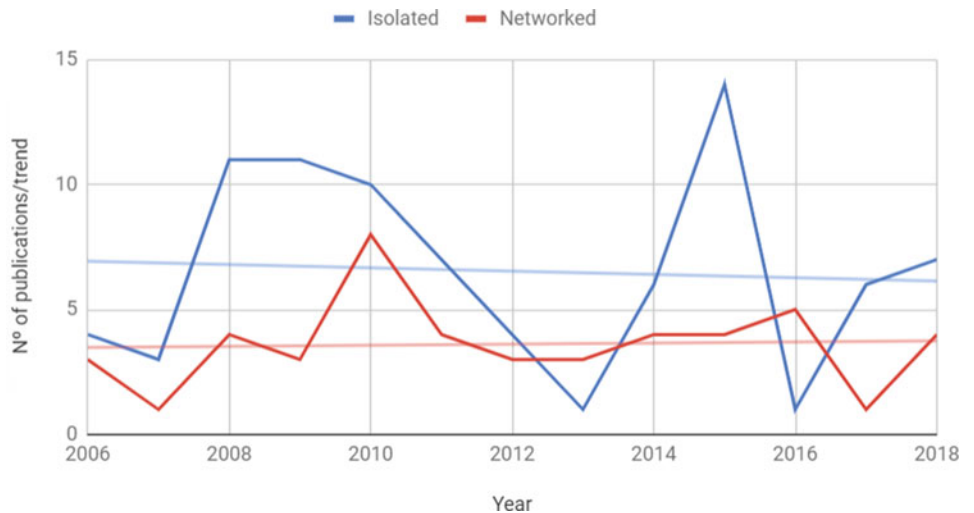
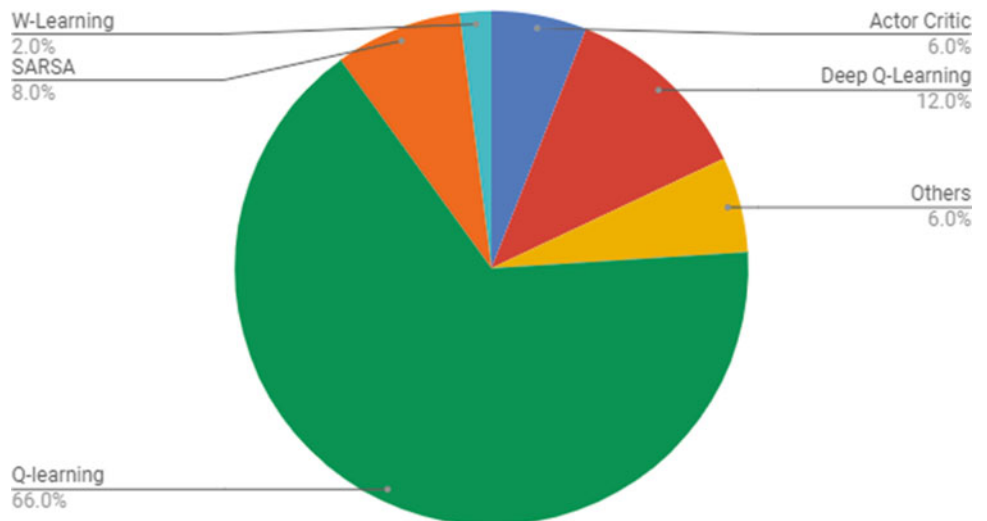


Fig. 2.6 Reinforcement learning approaches



between the processing power used in the simulation and research environments in relation to those in the equipment used for actual traffic management. Therefore, economic bias can make it impracticable to apply what has been studied in a real environment. It should also be considered that some studies propose models that require very specific tools and structures, that would result in an even higher implementation cost, such as [11], which suggests the use of Vehicle to Infrastructure networks (V2I), which requires that not only traffic lights, but also vehicles, have wireless devices capable of communicating with the network.

QC1: What Are the Trends and Gaps for This Field of Study? Efforts are mostly focused on the application of RL, especially Q-learning, without a clear definition of architecture. However, six studies were found that apply *Deep Q-learning*, an extension of the basic model. These publications are recent and date from 2016 and subsequent years—4 of them from 2018. Therefore, it is safe to say that there is a trend where more specialized and complex RL models, such as Deep Q-learning, have been gaining ground in the community.

Also, there is a lack of systems that apply Decision Trees, a branch of AI/ML, which, given an input set, performs cascade tests to finally deliver a decision consistent with the results obtained [15]. Since in vehicular traffic we have a set of indicators that, if analyzed together, make it possible to determine the current state and, consequently, the best decision for the system, it is believed that the referred technique could bring tangible solutions.

Another gap concerns the lack of application in real cases. Although some points have been raised as a justification, this gap creates a credibility problem. Although the field of study is well established and widely discussed, public agents and, mainly, the population, may still view it with suspicion. To mitigate this, it would be imperative to test these systems in real cases, with as much transparency as possible, to prove their efficiency and safety.

2.5 Discussion

As the number of publications addressing *Q-learning* stands out, with 33 articles out of 132 (25%), it is pertinent to investigate the reasons for this preference. One of its main characteristics is the adoption of an *online* learning model [3], which means that it can learn iteratively with the environment. This implies that, after the initial training phase, the model remains capable of learning. When considering that traffic behavior is intrinsically stochastic, it is notable that new situations, unknown during training, will appear, and the agent should be able to deal with them satisfactorily. Therefore, models that use *offline* learning algorithms, such as ANN, must be designed from the start with the ability

to deal with these situations or, at least, to circumvent them without significant performance losses.

In addition, *Q-learning* is *model-free* [19]; that is, it is not necessary to know the dynamics (transition probabilities) of the descriptive states of the process to learn and make decisions. Given the magnitude of the traffic control problem, fully describing its dynamics is virtually impossible, and techniques that need to know it beforehand, such as DP, should seek ways to compensate for this difficulty, which is also naturally compensated for in *Q-learning*.

When describing a signaled intersection in computational terms, one of the possible representations is through Markov Decision Processes (MDP) [2]. In MDP, each state is defined by a set of features that describe the environment, and the decision-making process takes the system from one state to another. Traffic lights can be easily described according to a MDP: the state can be defined by the color lit at the moment, the size of the queues, and the waiting time; and the system must decide at a given moment which way to be green-lit. *Q-learning* has its theoretical basis precisely in MDP [20]; therefore, its conformity with the management of a traffic light is inherent.

The nature of the problem makes *Q-learning* the most appropriate technique to tackle it. Even if other approaches emerge and present promising results, it is safe to say that there is a tendency for studies that combine techniques with *Q-learning* in order to improve it. For example, [17] compares the use of *Deep Q-Learning* with classic RL, increasing the number of descriptive variables while decreasing computational load and convergence time. Also, [22] combines *Recurrent Neural Networks* with *Deep Q-Learning*, in order to deal with previously unknown variables that affect the system, increasing the agent's generalization capacity.

2.6 Conclusion

An SMR was carried out, mapping the main AI and ML techniques used for vehicle traffic control. A total of 327 articles were found using three search engines: Scopus, IEEE Xplore, and ACM. Of these, 132 were selected, considered compatible with the inclusion criteria, and classified according to their techniques and structures.

It was found that RL is the most used ML model (36,9%), especially *Q-learning*, which represents 25% of the selected works. As *Q-learning* is an online learning model, model-free and MDP-based, its application in the studied context is natural and justifies its majority share. Recent works [17, 22] indicate that the area will continue to trend towards RL, seeking to combine its basic premise with modern approaches, in order to further improve its performance.

We did not notice an upward or downward trend in networked architectures, making it impossible for us to indicate

if it is going to dominate future studies. Also, it was impossible to find articles in which the theoretical model was tested in real scenarios, which is a research gap. However, with the natural evolution of techniques and computer systems, we hope that soon these agents will hit the streets and, then, we will be able to analyze them from this new perspective.

References

1. E.E. Agency, Size of the vehicle fleet in europe (2020). <https://www.eea.europa.eu/data-and-maps/indicators/size-of-the-vehicle-fleet/size-of-the-vehicle-fleet-10>
2. R. Bellman, A markovian decision process. *Indiana Univ. Math. J.* **6**, 679–684 (1957)
3. S. Ben-David, E. Kushilevitz, Y. Mansour, Online learning versus offline learning. *Machine Learning* **29**(1), 45–63 (1997)
4. M. Bielli, G. Ambrosino, M. Boero, M. Mastretta, Artificial intelligence techniques for urban traffic control. *Transp. Res. A Gen.* **25**(5) (1991)
5. E. Charniak, D. McDermott, *Introduction to Artificial Intelligence*. Addison-Wesley Series in Computer Science (Addison-Wesley, 1985)
6. DENATRAN: Frota de veículos (2018). www.denatran.gov.br/estatistica/237-frota-veiculos
7. I. Dusparic, J. Monteil, V. Cahill, Towards autonomic urban traffic control with collaborative multi-policy reinforcement learning, in *2016 IEEE 19th Int'l Conf. on Intelligent Transportation Systems (ITSC)* (IEEE, Rio de Janeiro, Brazil, 2016), pp. 2065–2070
8. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, USA, 2016)
9. S.O. Haykin, *Neural Networks and Learning Machines* (3rd edn.) (Pearson, Upper Saddle River, NJ, USA, 2008)
10. L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: A survey. *J. Artif. Intell. Res.* **4**, 237–285 (1996)
11. W. Liu, J. Liu, J. Peng, Z. Zhu, Cooperative multi-agent traffic signal control system using fast gradient-descent function approximation for v2i networks, in *2014 IEEE International Conf. on Communications (ICC)* (IEEE, Sydney, NSW, Australia, 2014), pp. 2562–2567
12. K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering, in *Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering* (Swindon, United Kingdom, 2008), pp. 68–77
13. J. Qiao, N. Yang, J. Gao, Two-stage fuzzy logic controller for signalized intersection. *IEEE Trans. Syst. Man Cybern. A Syst. Humans* **41**(1), 178–184 (2011)
14. S. Ravichandiran, *Hands-on Reinforcement Learning with Python* (Packt Publishing, Birmingham, United Kingdom, 2018)
15. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (3rd edn.) (Pearson, Edinburgh GateHarlow, United Kingdom, 2009)
16. A. Schwartz, A reinforcement learning method for maximizing undiscounted rewards, in *Proceedings of the Tenth International Conference on International Conference on Machine Learning* (Elsevier, San Francisco, CA, USA, 1993), pp. 298–305
17. S.M.A. Shabestary, B. Abdulhai, Deep learning vs. discrete reinforcement learning for adaptive traffic signal control, in *2018 21st Int'l Conf. on Intelligent Transportation Systems (ITSC)* (IEEE, Maui, HI, USA, 2018)
18. S. Smith, G. Barlow, X.F. Xie, Z. Rubinstein, Smart urban signal networks: Initial application of the surtrac adaptive traffic signal control system, in *Proc. of the Twenty-Third International Conf. on Automated Planning and Scheduling*, pp. 434–442, Rome, Italy (2013)
19. R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction* (A Bradford Book, 2018)
20. C.J.C.H. Watkins, P. Dayan, Q-learning. *Machine Learning* **8**(3–4), 279–292 (1992)
21. K.L.A. Yau, J. Qadir, H.L. Khoo, M.H. Ling, P. Komisarczuk, A survey on reinforcement learning models and algorithms for traffic signal control. *ACM Comput. Surv.* **50**(3), 1–38 (2017)
22. J. Zeng, J. Hu, Y. Zhang, Adaptive traffic signal control with deep recurrent q-learning, in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1215–1220 (IEEE, Changshu, China, 2018)
23. D. Zhao, Y. Dai, Z. Zhang, Computational intelligence in urban traffic signal control: A survey. *IEEE Trans. Syst. Man Cybern. C (Appl. Rev.)* **42**(4), 485–494 (2012)

Human-in-the-Loop Flight Training of a Quadcopter for Autonomous Systems

Luke Rogers and Alex Redei

Abstract

A software framework was developed connecting a Parrot AR 2.0 quadcopter to a full motion flight simulator at the Michigan Aerospace Center for Simulations. The combination of a drone with a flight simulator provides for precise remote operations without putting a human pilot at risk. We use the motion capabilities of our flight simulator to keep the pilot oriented consistently with the quadcopter. The result was a responsive system utilizing telemetry data to synchronize a flight simulator to a drone's movement with low latency. The proposed system was developed over the course of 30 weeks and put through its paces in our lab over two days. This paper outlines our methods, from the software architecture, to a detailed description of the hardware, to accomplish this aim and outlines some directions for further study.

Keywords

Drones · Flight simulators · Aviation · Drone racing · Human-in-the-loop · Autonomous systems · First person view · Path finding · Acrobatics · Flight maneuvering

3.1 Introduction

Flight simulation and drones are exiting new technologies applicable to many fields such as entertainment, search and rescue, defense, agriculture, and land surveys. A great deal of immersion is lost when a drone is piloted remotely using

a handheld controller, as the drone's orientation is different from the pilot's.

As drones grow in popularity, new ideas about how to use them surface. From Amazon using drones to deliver goods [1] to obstacle avoidance in drone racing [2], drones and their technologies are rapidly evolving.

The purpose of this project is to create a system that allows a user to fly a drone remotely from the inside of a flight simulator. The goal is to create an immersive experience that gives the pilot the feeling as though they were actually in the drone. This is a challenge as it requires low latency communication between the controls and the multiple computers used for computation and communication. With so many components working together in our system, low latency communication between the drone and flight simulator can be hard to achieve. Flight controllers and UAV's have been researched before through surveys of open-source drone platform elements, such as the Pixhawk [3], but here we extend this further.

From within the simulator, the pilot will view a live feed from the drone. This feed is taken from the forward-facing camera on the drone and will give the pilot a sense of first person point-of-view. The flight simulator being used is a 2-axis 360-degree capable flight simulator at the Michigan Aerospace Center for Simulations. The drone being used is a Parrot AR 2.0 drone. The key challenge was connecting two independent systems in a reliable and performant manner. For this, we have created a python script that both converts the input controls to data the drone can understand and sends the telemetry data from the drone to the flight simulator. Our python script is built off Phillip J. DeGraff's open source project that he created for the Parrot AR 2.0 drone. This code is designed as an SDK for the drone.

The challenge is not only to send the telemetry data from the drone to the flight simulator, but also in ensuring the data corresponds to movements that make sense in the simulator. Another challenge is latency in the video feed. A large latency between the simulator's movements and the video feed will

L. Rogers · A. Redei (✉)
Department of Computer Science, Central Michigan University,
Mt. Pleasant, MI, USA
e-mail: roger11t@cmich.edu; redei1a@cmich.edu

quickly result in motion sickness. Even a small delay would ruin the immersion.

Reliable connections between all the hardware components are also necessary for this project to succeed. Faulty or laggy connections will ensure data loss and result in less than satisfactory movement in the simulator. The quality and dependability of the software system is paramount here.

At the time of writing this paper, the system is not a finished product. However, the basic design of this software works. The simulator can move according to the movements of the drone. The video feed from the drone can be viewed, and a joystick implementation is being worked on now. Overall, the project is in a working state but is not fully polished, especially for end user performance standards.

The rest of this paper is organized as follows: Sect. 3.2 is devoted to the background and related work of other applications of drones, hardware is presented in Sect. 3.3, software architecture of this solution is presented in Sects. 3.4, 3.5 address the experimental results we gathered, and finally Sect. 3.6 wraps up our findings with several concluding remarks and directions for future work.

3.2 Related Work

There are three possible rotations any aerial object can perform. Expressed in Euler rotations, there is pitch, roll, and yaw. A NASA-made diagram demonstrating what each of these terms means for an aircraft is shown in Fig. 3.1.

These three rotations are the floating point values that are sent to our flight simulator. The code that we have written captures these 3 values and sends them via a UDP Packet to our controller PC. A .NET library called sim tools then converts this into actionable rotations that the simulator can use to move accordingly.

A current trend that leads us to believe that this technology will be very popular is drone racing. Drone racing is when “small unmanned aerial vehicles (UAVs) fly through a series of obstacles by human pilots” [5]. One of our ideas was to

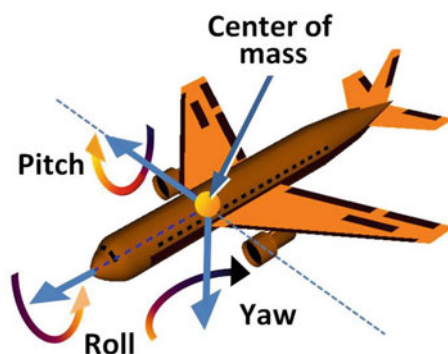


Fig. 3.1 Pitch, Roll, and Yaw as shown on an aircraft [4]

extend this into flight simulation. This will create a unique experience that takes an already popular hobby and adds a new dimension to it, similar to what is shown in Fig. 3.2.

In 2016, IEEE ran a contest where “the challenge consisted of developing a drone capable of racing a track autonomously” [6]. Naturally, this is different than our technology. While our system incorporates a human-in-the-loop, this highlights the immense popularity of drone racing. Many hobbyists are already looking at novel ways to make drone racing more fun. The framework described in this paper would adapt well to this kind of autonomous racing. In the event of a failure of the autonomous system, the drone can be easily recovered through manual control. A matter of fact, our system can do this right now if we enabled the autonomous capabilities of our Parrot AR 2.0 drone.

Still, drone racing is not the only way drones are being used. The military is highly invested in drones and autonomous systems. However, this raises the question of “Can artificial intelligence (AI) be more ethical than human intelligence?” [7]. While this paper is not about the ethics of drone use in the military, we want to respect this dilemma. The question becomes should drones in the military capable of killing be piloted by humans or not. This technology is designed in part with military use in mind. After all, “since 2001, the number of unmanned aerial vehicles (UAV’s) in the US military grew from 70 to 7000” [8]. If it is deemed that humans are more ethical than AI, then human pilots would be needed to audit autonomous systems. Our framework is designed for such situations. Our thought is this system can serve as either a training activity for future pilots, or as a special kind of tool to help keep drone pilots oriented.

Still, there are other fields where this system could be used. Drones have gained popularity in search and rescue missions. This often includes “be[ing] used to explore, for example, the remains after a catastrophe” [9]. One team is approaching drone rescue missions with AI. This team from Turkey is researching “a team of UAV’s deployed for searching for a single target, where the mission is considered successful once the target is found” [10]. Here, they developed an AI that was validated using Monte-Carlo simulations. If an autonomous system failed during the search and rescue mission, our system would be ready to allow a human pilots to take over and do many of those same things.

Others are using drones in more observatory roles. One team is using drones to map details in a forest canopy [11]. Drones are also being used to capture thermal images of urban land surface temperatures to help with urbanization and climate change [12].

Together, all of these different uses and work related to drones paints a very clear picture: drones are an integral part of our future. This framework creates a system that allows people to be more involved with drones in ways that are not widespread.

Fig. 3.2 An example of what a drone racing course may look like [13]

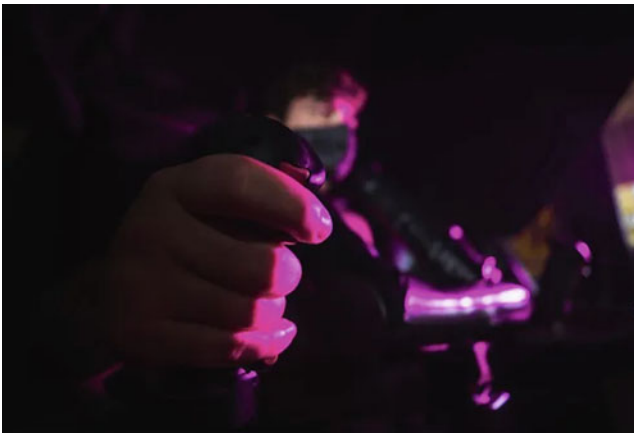
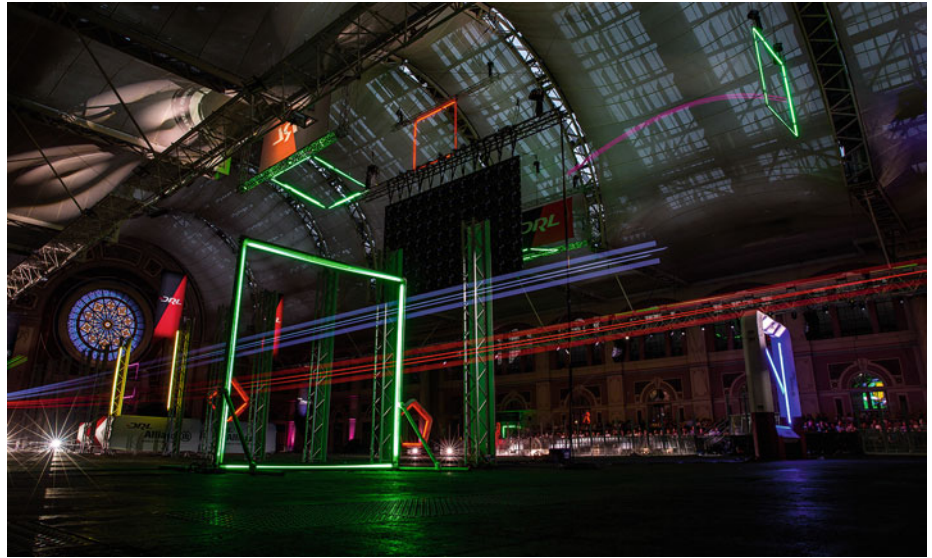


Fig. 3.3 Luke sitting inside the simulator

3.3 Hardware Used

For this work, we made use of our simulator. This simulator was provided by Dr. Alex Redei of the Computer Science Department at Central Michigan University. This simulator can move in pitch, and roll. This simulator comes fit with two sets of controls and chairs, as demonstrated by the author in Fig. 3.3. It also comes with speakers and a projector. For the sake of this project, only one set of controls and one chair will be used.

An important feature of the simulator is its ability stop for an emergency and be reset to its load position manually. In the case of a power outage or system failure, these features ensure the simulator is safe to use. In addition, the simulator has been approved by the Michigan government as safe to use.

In addition to the simulator, we use a Parrot AR Drone 2.0. This is a popular drone created by the company Parrot. It has also been used in other academic studies for activities such



Fig. 3.4 A picture of our first Parrot AR 2.0 Drone

as real-time object detection [14] This is a quadcopter drone. The drone is an older drone and as such, is not the quickest or most agile of drones. However, we chose to use this drone due to its ease of being programmed and how cheap it was. We knew testing this could result in many drone crashes, and it has, so having a replaceable drone was a must. The Parrot AR Drone 2.0 does face many challenges including a buggy video stream (due to its 2.4GHZ connection) and extreme drifting (due to battle-damage of many crashes). Both of these make it difficult to have precise controls, but do not prevent us from testing whether the core systems are working. The first Parrot AR Drone 2.0 we used can be seen below in Fig. 3.4.

We have a dedicated computer that runs our python scripts and communicates with the Parrot AR Drone 2.0. This runs on the Ubuntu 14 linux environment. The Python version on the computer must also be Python 2.7. This is due to the API we utilize from the code created by Philip J. DeGraf.

In addition to the dedicated computer, we have a controller PC. The controller PC runs Sim Tools which is the software that interprets the motion data and then sends it to the flight simulator. The dedicated PC and controller PC communicate via an Ethernet cable and UDP packets.

3.4 Software Design

First, our requirements for the system are listed below in Table 3.1.

A use case diagram is shown in Fig. 3.5. The use case diagram clarifies the interactions between the system, the pilot, the drone, and the flight simulator.

Many of the use cases are triggered via key presses. Some, such as send video, is automatically implemented into the system and a key press is not required. This is the case for the “Send Video” use case for the drone. The drone will automatically send video when the system is in startup.

The code is written in Python 2. It was designed in a text editor in a Linux environment. The code is then run from the command line. Our code utilizes Philip J. DeGraff’s API. His code can be found at www.playsheep.de.

The program begins once it is run from the command line. The dedicated PC then requests to connect to the drone. This is done via WiFi and the computer must be connected to the drone’s WiFi. Once this is done, the battery life is printed to the console, the video stream is displayed on the screen, and telemetry data is then sent.

The input is currently read via keyboard with the goal to implement a joystick in the Spring of 2021. The code listens for keyboard input so long as the while loop in which the program is running is not set to false. It is set to false by executing the shutdown, which is done with a key press of “P”. All actions for the drones are available via keyboard press. Some of these actions include but are not limited to, take off, land, move forward, turn, and move backward. For example, when the “W” key is pressed the drone will fly forward. When the “S” key is pressed the drone will fly backward. When “ ” is pressed the drone will takeoff if it is currently landed. If “ ” is pressed and the drone is flying it will land. Pressing “P” will end the program which causes the drone to land and the program to terminate. If “E” is pressed the drone will fly right. IF “7” is pressed the drone will turn -10 degrees.

In the event of a program failure, the simulator will stop moving. If the simulator does not receive any more telemetry

data, it simply stays in its last location. In the event the drone crashes, the simulator will move with the drone during the crash. However, once the drone detects it has crashed, it stops sending telemetry data. The simulator will then remain in that position. To let the pilots out of the simulator we can reset the position using our flight simulator software. This is a different program than what flies the drone. In the event of a power failure, the simulator has manual releases and the operator can manually move the simulator into place for the pilot to exit.

For future work we aim to have a joystick implemented into our system. This will work better with the controls already built into the flight simulator. Currently, one of the major challenges we face with this is implementing the joystick in Linux. We are using an Arduino board to map joystick movements to key presses based on Alex Redei’s work for multi-axis joystick inputs [15]. This works in Windows, but we are currently attempting to get it to work in Linux. This is still a work in progress. A screenshot of this system in Windows can be seen in Fig. 3.7.

3.5 Experimental Results

Due to COVID-19 restrictions our testing with human participants was highly limited. Much of the testing of the individual components of our system was done independently. There were two separate testing days where everything came together. On the first day, we were able to successfully move the simulator via the drone. This was tested by firing up our program, picking up the drone, and tilting it left, right, forward, upside down, and more. In each case the simulator moved accordingly. This can be seen below in Fig. 3.6.

A UDP packet is structured as a comma separated list with UTF-8 encoding containing each of the telmetry components. Using this data, the simulator moves according to the position of the drone. In addition to our handheld rotation test, we flew the drone in the lab and the simulator moved accordingly. During the first day of testing we encountered several bugs including the video stream did not appear, and the drone could only move in one direction forever.

On the second day of testing, we fixed the above two issues. First, we fixed the video. There was an else-if statement at the bottom of the video code that was causing the video feed to crash out. We removed this line and the video appeared. Second, we fixed the movement of the drone. Our else-if block failed to return the drone to a hovering state if no input was detected. Once this was added the drone could change direction. On this day we once again tested the simulator moving from outside of it. The simulator once again moved accordingly but this time with more precise movement and a video feed showing on screen.

Table 3.1 Software requirements

Requirement	Priority	Description
1	High	The software must be able to transmit telemetry data from the drone to the controller PC
2	High	The software must be able to move the simulator in correlation with the drone
3	High	The software must be able to display the video feed from the drone inside the simulator
4	High	The software must be able to control the movements of the drone

Fig. 3.5 The use case diagram

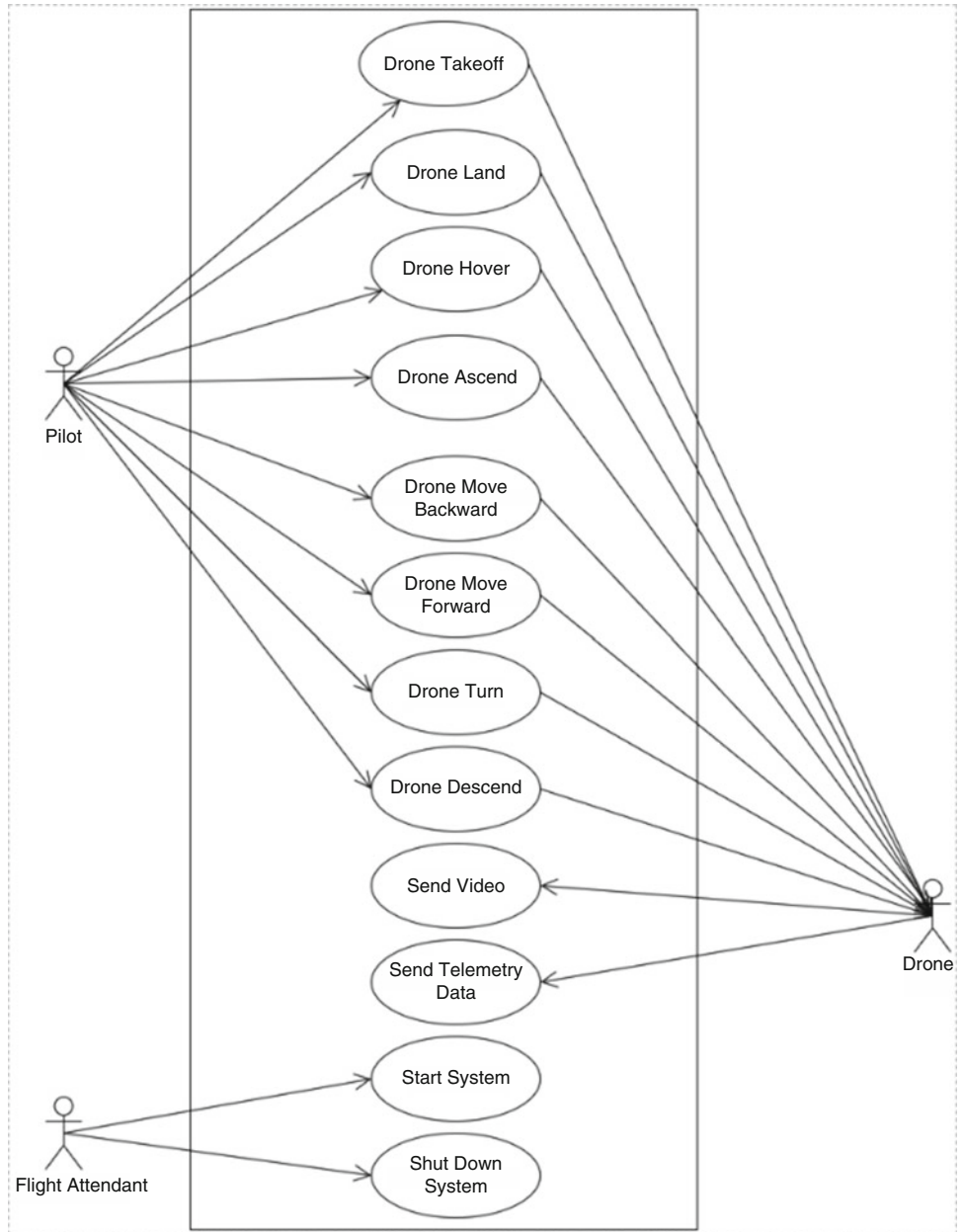


Fig. 3.6 Luke moving the drone and having the simulator move accordingly



Fig. 3.7 Our second day of Testing. (Left) Jack Fouch, (Middle) drone landed successfully (Right) telemetry data being printed to the console as it is being sent to the simulator



The results of our two tests show a very promising future for this system. The telemetry data was successfully sent from the drone to the dedicated PC. It was then rerouted to the controller PC and Sim Tools via a UDP packet. This data was then translated to motion for the simulator. The motion showed very low latency between the movements of the drone and the movements of the simulator (we were sending updates every 5/10,000 of a second). There also seemed to be negligible packet loss as the simulator did not seem to fall out of sync with the drone. All movements were smooth and correlated very closely to that of the drone.

As for the video feed, there was some observed latency, but we did not measure it. We tested this by running our hand in front of the drone camera and looking at the video feed at the same time. The pilot will rely on the visual feed for navigation. Due to COVID we have not yet tested with a real pilot in the simulator. We are unsure of the threshold at which latency will start to affect flight. Our belief, based on prior studies of human perception, is that as long as the simulator's motion and video feed are < 100 ms of latency that we should not have any significant issues with remote piloting of the drone.

An additional factor is the inertia of the simulator itself. The latency between the control input and the movement of the drone was present but was not major. When moving from a stopped position to a moving position the latency is low. When moving from a moving position to the opposite direction, such as front to back, the latency is much higher. This was not due to slow transmission of data, but the natural physics of stopping and changing direction in the air.

3.6 Conclusions and Future Work

In this paper, we outlined our framework for connecting a drone to a flight simulator. We examined the overall goal

for this system, some of its requirements, the general system structure, and the outcomes of our initial tests. Our initial testes are promising and indicates this system can be relevant in the entertainment, military, and emergency management fields. Thus far, our system is matching the performance attributes we anticipated as being needed for use in these fields.

Our system has successfully connected a drone to a flight simulator. Requirements 1–4 have all been met thus far in our research. With some minor adjustments this system can be accessible for end users soon.

Teams around the world involved in 5G and drones have also been looking at ways to reduce latency. One group is working on new scalable ways to improve connectivity [16]. In the future we hope to incorporate 5G connectivity to our drone, boosting the range and lowering the latency of the system.

Another potential avenue of future work is the incorporation of a future course prediction algorithm. The motions of the simulator is matched solely based on the drone's actual telemetry data, not a estimation of where it is headed. But since there is a hardware latency with the inertia's of the drone and simulator themselves, an improvement could incorporate machine learning to estimate where each system is heading in the near future. This would enhance the system's response time.

Acknowledgments We want to thank Matthew Drain, Jack Fouch, and Aaron Friedlein. All three worked on this project as part of their senior capstone for Computer Science at Central Michigan University. They provided help writing the code and designing many of the diagrams seen in this paper. Without them, this system would not be running like it is today,

Philip J. DeGraff also deserves a thank you. His code serves as the basis for our system. He makes his code available with tutorials at www.playsheep.de. Making his code accessible is a great service to our community and this project.

References

1. D. Sacramento, D. Pisinger, S. Ropke, An adaptive large neighborhood search metaheuristic for the vehicle routing problem with drones. *Transp. Res. C Emerg. Technol.* **102**, 289–315 (2019)
2. S.-Y. Shin, Y.-W. Kang, Y.-G. Kim, Obstacle avoidance drone by deep reinforcement learning and its racing with human pilot. *Applied Sciences* **9**(24), 5571 (2019)
3. E. Ebeid, M. Skriver, K.H. Terkildsen, K. Jensen, U.P. Schultz, A survey of open-source uav flight controllers and flight simulators. *Microprocess. Microsyst.* **61**, 11–20 (2018)
4. N. Hall, *Aircraft rotations* (2015)
5. S. Jung, S. Cho, D. Lee, H. Lee, D.H. Shim, A direct visual servoing based framework for the 2016 iros autonomous drone racing challenge, Aug 2017
6. L.O. Rojas-Perez, J. Martinez-Carranza, Deeppilot: A cmn for autonomous drone racing, Aug 2020
7. T. de Swarte, O. Boufous, P. Escalle, Artificial intelligence, ethics and human values: the cases of military drones and companion robots. *Artif. Life Robot.* **24**(3), 291–296 (2019)
8. P.M. Asaro, The labor of surveillance and bureaucratized killing: new subjectivities of military drone operators. *Social Semiotics* **23**(2), 196–224 (2013)
9. G.G. De la Torre, M.A. Ramallo, E. Cervantes, Workload perception in drone flight training simulators. *Comput. Human Behav.* **64**, 449–454 (2016)
10. E. Yanmaz Adam, Connectivity considerations for mission planning of a search and rescue drone team. *Turk. J. Electr. Eng. Comput. Sci.* **28**(4), 2228–2243 (2020)
11. J. Zhang, J. Hu, J. Lian, Z. Fan, X. Ouyang, W. Ye, Seeing the forest from drones: Testing the potential of lightweight drones as a tool for long-term forest monitoring. *Biological Conservation* **198**, 60–69 (2016)
12. J. Naughton, W. McDonald, Evaluating the variability of urban land surface temperatures using drone observations. *Remote Sensing (Basel, Switzerland)* **11**(14), 1722 (2019)
13. S. Paston, The drone racing league of london, 2017
14. A. Rohan, M. Rabah, S.-H. Kim, Convolutional neural network-based real-time object detection and tracking for parrot ar drone 2. *IEEE Access* **7**, 69575–69584 (2019)
15. A. Redei, S. Dascalu, A method for handling multi axis input for a motion based flight simulator, in *27th International Conference on Software and Data Engineering*, New Orleans, LA, 2018
16. P. Chandhar, E.G. Larsson, Massive mimo for connectivity with drones: Case studies and future directions. *IEEE Access* **7**, 94676–94691 (2019)

COVID-19: The Importance of Artificial Intelligence and Digital Health During a Pandemic

Maximilian Espuny, José S. da Motta Reis, Gabriel M. Monteiro Diogo, Thalita L. Reis Campos, Vitor H. de Mello Santos, Ana C. Ferreira Costa, Gildarcio S. Gonçalves, Paulo M. Tasinaffo, Luiz A. Vieira Dias, Adilson M. da Cunha, Nilo A. de Souza Sampaio, Andréia M. Rodrigues, and Otávio J. de Oliveira

Abstract

The Covid-19 has brought about a major change in the way people live, work and interact. To face the challenges of the epidemic, health professionals and researchers have implemented several technologies from Industry 4.0. In order to elucidate the application of these technologies in the context of the pandemic, the objective of this article is to analyze the main research trends of the Technologies 4.0 from the main publications on the subject. Data collection was carried out in the Scopus database in September 2020 and 413 studies were identified. The gaps identified in this research were: Apply artificial intelligence and I4.0 technologies to support and speed up Covid-19 diagnosis, Implement Risk Management tools to prevent and mitigate new Covid-19 infection waves, Integrate I4.0 technologies into microbiology and clinical trials, Mapping and sharing data that identify transmission rates and Covid' 19 diffusion routes, Search for treatment alternatives to Covid-19 through algorithms and artificial intelligence. The main academic contribution of this article was to systematize technological trends and under-

standing the influence of artificial intelligence and impact on the most urgent issues of the pandemic.

Keywords

Technologies · Industry 4.0 · Artificial intelligence · COVID-19 · SARS-CoV-2 · Digital health

4.1 Introduction

At the beginning of the twenty-first century, the SARS-CoV-1 epidemics occurred in 2002 and MERS-CoV in 2012. SARS-Cov-2 is the third disease from the same virus family in the past two decades [1]. The SARS-CoV-2 outbreak, which caused Covid-19, started in Wuhan in 2019 and soon reached a pandemic level with the statement by the World Health Organization (WHO) on 11 March 2020 [2].

Covid-19 has brought about a major change in the way people live, work, interact, requiring society to redesign living in public places and especially in closed environments, such as: homes, workplaces, public facilities and workplaces. Entertainment [3]. And to reduce coronavirus transmission, WHO and national disease control centers have issued several guidelines including social distance; frequent hand washing; and labels for sneezing and coughing, leaving the elbow flexed to contain the droplets of secretions [4].

To face the challenges of the epidemic, health professionals and researchers have implemented several technologies. Alternatives of a Deep Learning Model are being studied to detect Covid-19 in computed tomography scans and accurately distinguish possible cases of pneumonia [5]. The Taiwanese government has used the experience gained in combating Severe Acute Respiratory Syndrome in 2003 and using its Big Data to constantly inform the evolution of the pandemic and to respond in a precise and transpar-

M. Espuny · J. S. da Motta Reis · T. L. Reis Campos
V. H. de Mello Santos · A. C. Ferreira Costa · O. J. de Oliveira
São Paulo State University – UNESP, Production Department,
Guaratinguetá, Brazil

G. M. Monteiro Diogo · A. M. Rodrigues
São Paulo State University – UNESP, Administration Department,
Jaboticabal, Brazil

G. S. Gonçalves (✉) · P. M. Tasinaffo · L. A. Vieira Dias
A. M. da Cunha
Aeronautics Institute of Technology, ITA, Computer Science Division,
São José dos Campos, Brazil
e-mail: gildarcio@ita.br

N. A. de Souza Sampaio
State University of Rio de Janeiro, UERJ, Department of Mathematics,
Physics and Computing, Resende, Brazil

ent manner to its citizens [6]. Artificial Intelligence has been used constantly to improve the diagnosis of Covid-19, either through X-ray examinations or through computed tomography [7].

Although this research and others that make up this field of study have presented important perspectives, it is necessary to carry out more work to identify the contributions of advanced technologies to mitigate and solve the needs imposed by COVID-19. Given the above, the question that will guide this research is: what are the main paths taken by I4.0 technologies in the context of Covid-19? To answer it, the objective of this article is to analyze the main research trends of I4.0 technologies from the main publications on the subject.

In addition to this introduction section, the article will present the sections of the theoretical framework, research method, results, discussion, conclusion and references.

4.2 Theoretical Background

The new coronavirus (SARS-CoV-2), which causes COVID-19, is the cause of severe acute respiratory syndrome and transmission from person to person has caused it to spread rapidly in several countries since December 2019 [8]. Its symptoms may resemble those of a seasonal flu, but the potential diffuse alveolar damage means that in many cases its treatment requires advanced respiratory assistance, including artificial ventilation [9, 10].

COVID-19 has reached more than 3 million people and the speed of medical diagnosis is of great importance to enable agile isolation and treatment of infected patients, in order to prevent further spread of the virus [8]. In this context, information technologies such as Artificial Intelligence (AI) provide greater learning for doctors and have also been used to analyze the results of medical examinations and make the diagnosis faster and more accurate, while machine learning algorithms have been used to predict mortality and spread of the virus [5–7, 11, 12].

AI allows for agile new drug development and reorientation of existing drugs [13]. Thus, learning algorithms are used to identify drugs that are capable of inhibiting inflammation and infections caused by SARS-CoV-2 [14]. In addition, from the AI it is possible to perform a safer, more accurate and efficient image diagnosis to assist in the clinical and diagnostic evaluation of the patient, identify possible outbreaks and predict their propagation nature, thus helping to assist in decision making and definition of strategies related to coronavirus [7, 15].

4.3 Method

The research can be classified as basic, exploratory and qualitative approach. As a technical procedure, the literature review was adopted. Exploratory research aims to gain greater familiarity with a given phenomenon or gain new insights on the researched theme [16].

Data collection was carried out in the Scopus database in September 2020. The search searched only for articles and publications made in events or books were dispensed with. The following terms were used in the titles and keywords, all in the English language, in the search: “Artificial Intelligence” or “Data Science” or “Machine Learning” or “Blockchain” or “Big Data” or “Algorithms” or “Data Visualization” or “Data Analysis” and “COVID-19” or “SARS-CoV-2”.

In the research, 413 indexed studies were identified. For the identification of scientific gaps related to the research, the 20 articles most cited in the database were used, considering the time frame of 2020. The delimitation of 20 articles was carried out through non-statistical sampling, as it determines the significant impact on the conduct of the research [17]. The data were treated using Microsoft Excel software.

4.4 Results

The 20 most relevant articles are presented, sorted from the most to the least cited documents, according to Table 4.1. All obtained at least 5 citations and 13 of the 20 articles had at least 10 or more citations, which may indicate that despite having been published in a short time, they have influenced the Covid-19 theme associated with the most sophisticated technologies in 2020.

According to Table 4.2, research gaps involving the main themes of Covid-19 and Technology were grouped, totaling five groups.

In “Apply artificial intelligence and I4.0 technologies to support and speed up Covid-19 diagnosis” the study opportunities are mentioned that propose application of algorithms to increase the efficiency in the detection of the new coronavirus. The algorithmic solution is suggested mainly in the X-ray and computed tomography exams in the chest [30]. Both tests have been shown to be more agile than RT-PCR, which although it is a test intended for a specific purpose and should not be missed in the detection of Covid-19 can take up to 2 days to complete [8, 20]. This procedure is important for emergency cases that may suggest immediate hospitalization.

Table 4.1 Scientific gaps of the 20 most mentioned works

Title	Authors	Source	Citation	Scientific gaps
Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing	Wang et al. (2020)	Journal of the American Medical Association	162	Identify the effectiveness of using smart technologies to mitigate the impacts of new pandemics
Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy	Li et al. (2020)	Radiology	53	Assess the ability of deep learning to diagnose COVID-19
Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis	Pirouz et al. (2020)	Sustainability	26	Identify the impacts of COVID-19 on social sustainability
Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19	Shi et al. (2020)	IEEE Reviews in Biomedical Engineering	20	Identify the effectiveness of artificial intelligence in the diagnosis of COVID-19
Artificial intelligence (AI) applications for COVID-19 pandemic	Vaishya et al. (2020)	Diabetes and Metabolic Syndrome: Clinical Research and Reviews	19	Analyze the effectiveness of artificial intelligence in preventing COVID-19
Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm	Wang et al. (2020)	Science of the Total Environment	18	Use patient information based algorithm to estimate the spread rate of a virus in real time
Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine	Srinivasa et al. (2020)	Infection Control and Hospital Epidemiology	16	Evaluate the effectiveness of machine learning in the identification of COVID-19 cases
First data analysis about possible COVID-19 virus airborne diffusion due to air particulate matter (PM): the case of Lombardy (Italy)	Bontempi et al. (2020)	Environmental Research	15	Propose the use of artificial intelligence to identify the broadcast and transmission routes of COVID-19
A British Society of Thoracic Imaging statement: considerations in designing local imaging diagnostic algorithms for the COVID-19 pandemic	Nair et al. (2020)	Clinical Radiology	13	Identify the effectiveness of using algorithms in COVID-19 imaging diagnosis
Analyzing the epidemiological outbreak of COVID-19: a visual exploratory data analysis approach	Dey et al. (2020)	Journal of Medical Virology	11	Propose the implementation of technologies to mitigate new waves of COVID-19 infection
Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study	Randhawa et al. (2020)	PLoS One	11	Evaluate the feasibility of using artificial intelligence for genomic classification of the COVID-19 virus in time
Artificial intelligence and machine learning to fight covid-19	Alimadadi t al. (2020)	Physiological Genomics	11	Propose the use of smart technologies that facilitate access to COVID-19 data
Integrated radiologic algorithm for COVID-19 pandemic	Sverzellati et al. (2020)	Journal of Thoracic Imaging	10	To assess the accuracy of the prognosis integrated radiological algorithm for screening patients with suspected COVID-19
Artificial intelligence-enabled rapid diagnosis of patients with COVID-19	Mei et al. (2020)	Nature Medicine	9	Identify the benefits of artificial intelligence to diagnose patients with COVID-19 through computed tomography analysis

(continued)

Table 4.1 (continued)

Title	Authors	Source	Citation	Scientific gaps
Laparoscopy at all costs? Not now during COVID-19 outbreak and not for acute care surgery and emergency colorectal surgery: a practical algorithm from a hub tertiary teaching hospital in northern Lombardy, Italy	Saverio et al. (2020)	Journal of Trauma and Acute Care Surgery	9	Assess which surgical method is safer during a pandemic: laparoscopy or laparotomy
An algorithm for managing QT prolongation in coronavirus disease 2019 (COVID-19) patients treated with either chloroquine or hydroxychloroquine in conjunction with azithromycin: possible benefits of intravenous lidocaine	Mitra et al. (2020)	HeartRhythm Case Reports	9	Investigate whether this algorithm can be applied to patients positive for COVID-19 who need this therapy
Mechanism of baricitinib supports artificial intelligence-predicted testing in COVID-19 patients	Stebbing et al. (2020)	EMBO Molecular Medicine	8	To evaluate the efficiency and safety of treatment with baricitinib in randomized clinical trials
Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 patients	Yun et al. (2020)	Clinica Chimica Acta	8	Assess whether there is competition between COVID-19 and Influenza viruses and what is the relationship between eosinophils and the severity of COVID-19 infection
Use of CT and artificial intelligence in suspected or COVID-19 positive patients: statement of the Italian Society of Medical and Interventional Radiology	Neri et al. (2020)	Medical Radiology	7	Validate the efficiency of using artificial intelligence as an auxiliary method of diagnosis and prognosis for patients with COVID-19
Artificial intelligence approach fighting COVID-19 with repurposing drugs	Ke et al. (2020)	Biomedical Journal	7	To evaluate the efficiency against COVID-19 of the drugs suggested by artificial intelligence in vitro and in vivo

Table 4.2 Scientific gap groups

Macro grouping	Scientific gaps
Apply artificial intelligence and I4.0 technologies to support and speed up Covid-19 diagnosis	Li et al. (2020); Mei et al. (2020); Nair et al. (2020); Neri et al. (2020); Shi et al. (2020); Srinivasa Rao and Vazquez (2020); Sverzellati et al. (2020).
Implement Risk Management tools to prevent and mitigate new Covid-19 infection waves	Dey et al. (2020); Di Saverio et al. (2020); Pirouz et al. (2020); Vaishya et al. (2020); Wang et al. (2020).
Integrate I4.0 technologies into microbiology and clinical trials	Ke et al. (2020); Randhawa et al. (2020); Yun et al. (2020).
Mapping and sharing data that identify transmission rates and COVID-19 diffusion routes	Alimadadi et al. (2020); Bontempi (2020); L. Wang et al. (2020).
Search for treatment alternatives to Covid-19 through algorithms and artificial intelligence	Mitra et al. (2020); Stebbing et al. (2020).

In the “Implement Risk Management tools to prevent and mitigate new Covid-19 infection waves” cluster, research opportunities are based on resources that can mitigate the effects of the pandemic on society. The possibility of using the Group Method of Data Handling (GMDH) algorithm to understand the correlations of average temperatures and humidity in the number of infections in Covid-19 is mentioned [21]. Another important opportunity for the prevention of pandemics is a database that understands and integrates information from various locations in the world, so that the

development of pandemics can be tracked in real time and thus stop its spread more quickly [22].

Regarding the “Integrate I4.0 technologies into microbiology and clinical trials” cluster, research opportunities are presented around the application of the method combines supervised Machine Learning, improving the quality of teaching learning and consequently the quality of life. Digital Signal Processing (MLDSP), providing analyzes of more than 5000 unique viral sequences [24, 31]. Studies are also indicated to improve the diagnoses that can distinguish Covid-19 from Influenza A/B, mainly be-

cause the research indicates that despite the outbreak of the new coronavirus, contamination by Influenza A/B is still greater and that it is difficult to diagnose a patient with both pathologies [25].

In “Mapping and sharing data that identify transmission rates and COVID’19 diffusion routes” one of the proposed tools is the use of Patient Information Based Algorithm (PIBA) to estimate in real time the mortality rate provided by the outbreak of the new coronavirus [6]. Studies are proposed that seek to measure the correlation between air pollution and the transmissibility of the new coronavirus, with the particulate matter (PM) being possibly responsible for the diffusion [27]. Scientists working in the fields of artificial intelligence and machine learning have been looking for ways to identify data from infected by Covid-10 through physiological characteristics, involuntary body gestures and therapeutic results [28].

In the “Search for treatment alternatives to Covid-19 through algorithms and artificial intelligence” cluster, the use of AI is proposed to identify old drugs that may be efficient for the treatment of SARS-CoV-2 and for the identification of more specific drugs, which they allow a more in-depth analysis of those who obtained a satisfactory result in small samples [13, 14, 29]. It is worth remembering that in the case of treatment options against the virus, there is a speed of studies that is very intensified and that changes constantly [14].

4.5 Conclusion

The objective of the work to analyze the main research trends of I4.0 technologies from the main publications on the subject was duly achieved.

The main academic contribution of this article was to systematize trends, allowing a better understanding of how I4.0 technologies are impacting the most urgent issues of the pandemic.

The most important applied contribution was to promote the solutions developed by hospitals, health professionals and the entire scientific community, so that the governments of the municipal, regional and federal levels can promote or even invest to fight the Covid-19 outbreak, as well as obtain the know-how for possible future pandemics.

It is desirable that the results of this research can reach the public, to make them aware of the efforts made to combat the pandemic, employing immeasurable technological and financial human resources. The dissemination of this information can help the various public health stakeholders, economic agents and public authorities to combat the ignorance of groups such as “antivaccines” and even those who completely despise the severity of the new coronavirus.

The main limitation of studies has been the linear number of information that has impacted on a constant change from the current recommendations, mainly because of the resources employed, as it has not been seen in the scientific and health community for a long time.

As a proposal for future studies, bibliographic reviews or even bibliometric studies are recommended that delimit the various technologies of I4.0 applied to Covid-19, to catalog and measure scientific production and serve as possible theoretical structuring scripts to be replicated in other potentials pandemics.

Acknowledgments This work was made possible with the support of CNPq Proc. 312894/2017-1. We thank CAPES for the financial support, The ITA - the Brazilian Aeronautics Institute of Technology and State Technology Educational Center Paula Souza – CPS.

References

1. B.N. Kulkarni, V. Anantharama, Repercussions of COVID-19 pandemic on municipal solid waste management: Challenges and opportunities. *Sci. Total Environ.* **743**, 140693 (2020). <https://doi.org/10.1016/j.scitotenv.2020.140693>
2. F. Di Maria, E. Beccaloni, L. Bonadonna, C. Cini, E. Confalonieri, G. La Rosa, M.R. Milana, E. Testai, F. Scaini, Minimization of spreading of SARS-CoV-2 via household waste produced by subjects affected by COVID-19 or in quarantine. *Sci. Total Environ.* **743**. Elsevier B.V., 140803 (2020). <https://doi.org/10.1016/j.scitotenv.2020.140803>
3. D. D’alessandro, M. Gola, L. Appolloni, M. Dettori, G.M. Fara, A. Rebecchi, G. Settimo, S. Capolongo, COVID-19 and living space challenge. Well-being and public health recommendations for a healthy, safe, and sustainable housing. *Acta Biomed* **91**, 61–75 (2020). <https://doi.org/10.23750/abm.v91i9-S.10115>
4. C. Nzediegwu, S.X. Chang, Improper solid waste management increases potential for COVID-19 spread in developing countries. *Resour. Conserv. Recycl.* **161**. Elsevier, 104947 (2020). <https://doi.org/10.1016/j.resconrec.2020.104947>
5. L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, et al., Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy. *Radiology* **296**, E65–E71 (2020). <https://doi.org/10.1148/radiol.2020200905>
6. C.J. Wang, C.Y. Ng, R.H. Brook, Response to COVID-19 in Taiwan: Big data analytics, new technology, and proactive testing. *JAMA* **323**, 1341 (2020). <https://doi.org/10.1001/jama.2020.3151>
7. F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* **3333**, 1–1 (2020). <https://doi.org/10.1109/RBME.2020.2987975>
8. X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, et al., Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **26**. Springer US, 1224–1228 (2020). <https://doi.org/10.1038/s41591-020-0931-3>
9. N. Sverzellati, G. Milanese, F. Milone, M. Balbi, R.E. Ledda, M. Silva, Integrated radiologic algorithm for COVID-19 pandemic. *J. Thorac. Imaging* **35**, 228–233 (2020). <https://doi.org/10.1097/RTI.0000000000000516>
10. J.S.d.M. Reis, F.D.O. Silva, M. Espuny, L.G.L. Alexandre, L.C.F.M. Barbosa, G. Santos, A.C.M. Bonassa, A.M. Faria,

- N.A.d.S. Sampaio, O.J. de Oliveira, The rapid escalation of publications on Covid-19: a snapshot of trends in the early months to overcome the pandemic and to improve life quality. *Int. J. Qual. Res.* **14**, 951–968 (2020). <https://doi.org/10.24874/IJQR14.03-19>
11. R. Vaishya, M. Javaid, I.H. Khan, A. Haleem, Artificial intelligence (AI) applications for COVID-19 pandemic. *Diabetes Metab. Syndr. Clin. Res. Rev.* **14**, 337–339 (2020). <https://doi.org/10.1016/j.dsx.2020.04.012>
 12. J.S.d.M. Reis, A.C.F. Costa, M. Espuny, W.J. Batista, F.E. Francisco, G.S. Gonçalves, P.M. Tasinaffo, L.A.V. Dias, A.M. da Cunha, O.J. de Oliveira, Education 4.0: Gaps research between school formation and technological development, in *17th International Conference on Information Technology–New Generations (ITNG 2020)*, ed. by S. Latifi, 1st edn., (Springer, Cham, 2020), pp. 415–420. https://doi.org/10.1007/978-3-030-43020-7_55
 13. Y.-Y. Ke, T.-T. Peng, T.-K. Yeh, W.-Z. Huang, S.-E. Chang, S.-H. Wu, H.-C. Hung, et al., Artificial intelligence approach fighting COVID-19 with repurposing drugs. *Biom. J.* **43**, 355–362 (2020). <https://doi.org/10.1016/j.bj.2020.05.001>
 14. J. Stebbing, V. Krishnan, S. Bono, S. Ottaviani, G. Casalini, P.J. Richardson, V. Monteil, et al., Mechanism of baricitinib supports artificial intelligence-predicted testing in COVID-19 patients. *EMBO Mol. Med.* **12**, 1–15 (2020). <https://doi.org/10.15252/emmm.202012697>
 15. K.C. Santosh, AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multidimensional/multimodal data. *J. Med. Syst.* **44**, 93 (2020). <https://doi.org/10.1007/s10916-020-01562-1>
 16. C.R. Kothari, G. Garg, *Research Methodology Methods and Techniques*, 4th edn. (New Age International, Nova Deli, 2019)
 17. D.P. Van Der Nest, L. Smidt, D. Lubbe, The application of statistical and/or non-statistical sampling techniques by internal audit functions in the South African banking industry. *Risk Gov. Control Financ. Mark. Inst.* **5**, 71–80 (2015). <https://doi.org/10.22495/rgcv5i1art7>
 18. S. Rao, S.R. Arni, J.A. Vazquez, Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. *Infect. Control Hosp. Epidemiol.* **41**, 826–830 (2020). <https://doi.org/10.1017/ice.2020.61>
 19. A. Nair, J.C.L. Rodrigues, S. Hare, A. Edey, A. Devaraj, J. Jacob, A. Johnstone, R. McStay, E. Denton, G. Robinson, A British Society of Thoracic Imaging statement: Considerations in designing local imaging diagnostic algorithms for the COVID-19 pandemic. *Clin. Radiol.* **75**. Elsevier Ltd, 329–334 (2020). <https://doi.org/10.1016/j.crad.2020.03.008>
 20. E. Neri, V. Miele, F. Coppola, R. Grassi, Use of CT and artificial intelligence in suspected or COVID-19 positive patients: Statement of the Italian Society of Medical and Interventional Radiology. *Radiol. Med.* **125**. Springer Milan, 505–508 (2020). <https://doi.org/10.1007/s11547-020-01197-9>
 21. B. Pirouz, S.S. Haghshenas, S.S. Haghshenas, P. Piro, Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis. *Sustainability* **12**, 2427 (2020). <https://doi.org/10.3390/su12062427>
 22. S.K. Dey, M.M. Rahman, U.R. Siddiqi, A. Howlader, Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. *J. Med. Virol.* **92**, 632–638 (2020). <https://doi.org/10.1002/jmv.25743>
 23. S. Di Saverio, M. Khan, F. Pata, G. Ietto, B. De Simone, E. Zani, G. Carcano, Laparoscopy at all costs? Not now during COVID-19 outbreak and not for acute care surgery and emergency colorectal surgery: A practical algorithm from a hub tertiary teaching hospital in Northern Lombardy, Italy. *J. Trauma Acute Care Surg.* **88**, 715–718 (2020). <https://doi.org/10.1097/TA.0000000000002727>
 24. G.S. Randhawa, M.P.M. Soltysiak, H. El Roz, C.P.E. de Souza, K.A. Hill, L. Kari, Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. Edited by Oliver Schildgen. *PLoS One* **15**, e0232391 (2020). <https://doi.org/10.1371/journal.pone.0232391>
 25. H. Yun, Z. Sun, J. Wu, A. Tang, M. Hu, Z. Xiang, Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 patients. *Clin. Chim. Acta* **507**. Elsevier, 94–97 (2020). <https://doi.org/10.1016/j.cca.2020.04.018>
 26. L. Wang, J. Li, S. Guo, N. Xie, L. Yao, Y. Cao, S.W. Day, et al., Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. *Sci. Total Environ.* **727**. Elsevier B.V., 138394 (2020). <https://doi.org/10.1016/j.scitotenv.2020.138394>
 27. E. Bontempi, First data analysis about possible COVID-19 virus airborne diffusion due to air particulate matter (PM): The case of Lombardy (Italy). *Environ. Res.* **186**. Elsevier Inc., 109639 (2020). <https://doi.org/10.1016/j.envres.2020.109639>
 28. A. Alimadadi, S. Aryal, I. Manandhar, P.B. Munroe, B. Joe, X. Cheng, Artificial intelligence and machine learning to fight COVID-19. *Physiol. Genomics* **52**, 200–202 (2020). <https://doi.org/10.1152/physiolgenomics.00029.2020>
 29. R.L. Mitra, S.A. Greenstein, L.M. Epstein, An algorithm for managing QT prolongation in coronavirus disease 2019 (COVID-19) patients treated with either chloroquine or hydroxychloroquine in conjunction with azithromycin: Possible benefits of intravenous lidocaine. *HeartRhythm Case Rep.* **6**. Elsevier Inc., 244–248 (2020). <https://doi.org/10.1016/j.hrcr.2020.03.016>
 30. H. Chen, J. Guo, C. Wang, F. Luo, X. Yu, W. Zhang, J. Li, et al., Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records. *Lancet* **395**, 809–815 (2020). [https://doi.org/10.1016/S0140-6736\(20\)30360-3](https://doi.org/10.1016/S0140-6736(20)30360-3)
 31. A.B.C.d.S. Alvarenga, M. Espuny, J.S.d.M. Reis, F.D.O. Silva, N.A.d.S. Sampaio, T.V. Nunhes, L.C.F.M. Barbosa, G. Santos, O.J. de Oliveira, The main perspectives of the quality of life of students in the secondary cycle: An overview of the opportunities, challenges and their greatest impact elements. *Int. J. Qual. Res.*, 15 (2021). <https://doi.org/10.24874/IJQR15.03-19>

CropWaterNeed: A Machine Learning Approach for Smart Agriculture

Malek Fredj, Rima Grati, and Khouloud Boukadi

Abstract

In this paper, we propose an approach CropWaterNeed in order to estimate and predict the future water needs and maximize the productivity in the irrigated areas. Unfortunately, we have not identified data available to be employed in such machine learning process in order to predict plants water needs. The proposed approach consists of extending the classic machine learning process. Particularly, we define a process to build dataset that contains plant water requirements. To collect data, we extract meteorological data from Climwat database and plants water requirements using Cropwat Tool. Then, we aggregate the extracted data into a dataset. Subsequently, we use the dataset to perform the learning process using XGBRegressor, Decision Tree, Random Forest and Gradient Boost Regressor. Afterward, we evaluate the model generated by each algorithm by measuring the performance measures such as MSE, RMSE and MAE. Our work shows that the model generated by XGBRegressor is the most efficient in our case while Random Forest is the least efficient. As future work, we aim to apply the proposed process to test the performance of other regression algorithms and to test the impact of using deep learning techniques with the extracted data.

Keywords

Artificial intelligence · Agriculture · Algorithms · Machine learning · Smart regression · Irrigation ·

M. Fredj (✉) · K. Boukadi
Mir@cl Laboratory, University of Sfax, Sfax, Tunisia
e-mail: khouloud.boukadi@fsegs.usf.tn

R. Grati
Zayed University, Abu Dhabi, United Arab Emirates
e-mail: rима.grati@zu.ac.ae

Supervised learning · CropWat · ClimWat · Irrigation requirement

5.1 Introduction

Agriculture plays a critical role in the global economy. The origin of agriculture goes back about ten thousand years, or about four hundred human generations in prehistoric times. Today, the increase in the world's population has caused the needs for food. As a result, the traditional methods of farmers become insufficient. In order to overcome the limitations of traditional methods, smart agriculture has emerged. It aims to automate, reduce costs and improve the quality of agricultural production. This new paradigm brings intelligence into traditional farming methods such as smart irrigation systems. These refer to the combination of the hardware devices and software applications with different technologies including among others machine learning. In general, machine learning has emerged with large data technologies and high-performance computing to create new opportunities for data science in the multidisciplinary field of agriculture technologies. Among these disciplines: crop management, including applications on forecasting yields, disease detection, weed detection and quality crops as well as water and soil management [1]. To do this, machine learning algorithms require a large amount of data to analyze crop performance in different climates. Based on these data, they are able to build a model that could predict the need of plants in terms of water.

In this context fits our research work which is also a part of the PRECIMED project.¹ This is an European project accepted under the PRIMA call, which deals with smart

¹www.precimed.eu

agriculture. It consists of optimizing the sustainability of agriculture using new water management practices. As part of this project, the objective of this work consists in processing agriculture-related data and developing a prediction model of the crops irrigation needs.

In the literature, some researches have used machine learning algorithms such as Support Vector Regression [2], Forest Regression [3], Recurrent Neural Networks [4], and Extreme Learning Machine [5] for water quantity prediction as well as the weekly evapotranspiration. Despite the variety of the proposed algorithms, the different contributions do not support the diversity of cultures and the availability of data that remains a major problem. To try to meet the limitations of the state of the art, we propose the approach CropWaterNeed that consists of taking advantage of the classic process of machine learning in order to determine the plant's needs in terms of water while considering the weather observations, soil parameters and plant needs in terms of water. CropWaterNeed's merit is its ability to examine different parameters involved in the prediction of water needs. These parameters were carefully identified following an extensive literature review exercise. In addition, CropWaterNeed's considers a variety of crops such as tomatoes, grapes and lemon for prediction water requirements for plants with a development cycle, a quantity of water to irrigate and a different cycle duration.

The rest of this paper is organized as follows: Section 5.2 presents the background that constitutes the basic of this work. Section 5.3 presents the literature review. Section 5.4 discusses our contribution CropWaterNeed for improving the process of machine learning and adapting it to smart agriculture to predict the amount of water to irrigate. Section 5.5 discusses the experimental results. Finally, Sect. 5.6 draws several conclusions.

5.2 Background

This section briefly introduces three concepts that form the basis of our work: Smart Agriculture, Machine learning and Big data in Smart Agriculture.

5.2.1 Smart Agriculture

Smart Agriculture is an emerging concept that helps the farmers to manage farms and make proper strategic and operational decisions using technologies such as IoT, drones, robots and AI. This enhances the quantity and quality of products while optimizing the human labor needed by production. Several agricultural applications are currently implemented using the Internet of Things (IoT) such as: Irrigation management system, pest and disease control, monitor-

ing of livestock movements, Monitoring of dairy products, Water quality monitoring, Greenhouse condition monitoring, Precision agriculture by drone and agricultural logistics management.

5.2.2 Machine Learning

Machine learning is an application of artificial intelligence (AI) that gives systems the ability to automatically learn and develop from experience without being explicitly programmed [6].

Essentially, through data and experience exposure, ML is the capacity of software or a computer to enhance the efficiency of tasks. First, a typical ML model learns the knowledge from the data to which it is exposed and then uses this knowledge to provide predictions of emerging future data. Machine learning aims to solve a number of problems among which we distinguish the following two problems: classification and regression problems. The first consists of predicting the value of a discrete variable that can take only certain values. While the second predicts the value of a continuous variable that can take an infinity of values.

In this research work, we aim to deal with the regression problem because we are trying to predict the amount of water.

There are several machine learning algorithms that deal with smart agriculture, we present in what follows the most used in the literature:

- K Nearest Neighbors (KNN) [7]: it is a simple algorithm that stores all available cases and classifies new cases based on a similarity measurement (for example, distance functions). K nearest neighbors has been used in statistical estimation and pattern recognition.
- Random Forest: It consists of a set of tree predictors, and each of them is constructed using an injection of chance [8]
- Decision Tree: It is mainly a method of building whole decision rules on predictor variables. Unlike the techniques of classical regression where the relation between the response and the predictors is predefined (e.g. straight line, quadratic), the Decision Tree provides the possibility of interactions and non-linearities between variables [9].
- Gradient Boost Regressor: Amplification, or amplified regression, is a recent data mining technique that has shown considerable success in predictive precision [10]. The Gradient Boost modifies the sample (by placing the labels on the negative gradient) while maintaining the distribution constant [11].
- XGBoost Regressor: it is the abbreviation for eXtreme Gradient Boosting. It is an efficient and scalable implementation of the gradient boosting framework proposed by [12, 13]

5.3 Related Work

To circumvent challenges such as crop diversity and large amounts of data, many researches have used machine learning algorithms to optimize the sustainability of agriculture. We present in this section relevant works that dealt with prediction systems in agriculture.

Vij et al. [3] highlighted the problems associated with rising sea levels and the need to have appropriate methods for maintaining crops that aim to reduce minimum excessive waste of water. To do this, the authors use the machine learning algorithms to predict irrigation patterns based on crops and weather scenarios. The solution suggested by the authors is based on the use of IoT (Internet of Things) and on an wireless sensor network installed in the farm. This network includes sensors, transducers, and actuators to monitor and control soil temperature, moisture, and fertility. In the network wireless sensors, each node is interconnected by a Wi-Fi module and transmits the data on a common server. The latter uses an automated python script to query data and send an alert or start signal for the required operation. In order to predict irrigation intervals, this system uses meteorological data which will be permanently from online open source API. In case of probability of rainfall greater than 98%, the target field will not be irrigated. However, for safety reasons if the humidity drops below a certain threshold for a specific crop, the field will be irrigated.

Salim and Mitton [14] presented an algorithm for reducing the amount of data transmitted between the sensors and the base station using automatic learning techniques. This algorithm mainly focuses on environmental data for the benefit of agriculture. Reducing the amount of data transmitted reduces energy consumption and bandwidth usage while maintaining the accuracy of the information. This approach requires the availability of different correlated data and mainly focuses on the amount of data transmitted not on the amount of data entered. For the implementation of this algorithm, the authors used a MATLAB simulator with a dataset of temperature in Lille, France from the Weather Underground website which collects data from a network of sensors made up of different stations.

In the same context, Adeyemi et al. [4] presented a dynamic neural network approach for modeling the temporal soil moisture fluxes. The models designed by this approach have generated robust predictions of soil for independent sites that were not used during the learning phase. The application of dynamic neural network models in a programming system predictive irrigation has been demonstrated using AQUACROP simulations of the growing season of potatoes. Other work [5] has looked at the evapotranspiration of crops as a basis for the determination of irrigation. Patil et al. [5] used the Extreme algorithm Learning Machine (ELM)

to estimate the weekly reference evapotranspiration (ET_o). This study evaluated the performance of three different input combinations, the first input combination used consists of the relative data at maximum and minimum air temperatures, while the second and third combination use the ET_o values of another station (extrinsic inputs) as well as the temperature data available locally as inputs.

Mousa et al. [15] proposed an approach to forecast the necessary irrigation intervals and the irrigation time applied. The proposed approach encompasses four tasks, namely the estimation of evapotranspiration, soil moisture monitoring, estimating of irrigation amount needed and calculation of irrigation time and irrigation schedule. The approach predicts the water needs of crops and the amount of water required based on weather conditions and soil moisture. These conditions are collected by sensors distributed throughout the concerned farm. Thus, they achieved simulations using MATLAB software.

The analysis of the related work reveals that there are several research studies that address smart agriculture based on machine learning. However, despite the diversity of work, the availability of data remains a major problem and no complete data source has not been provided in the literature. In addition, there are data that are insufficiently sized for the accuracy of the automatic learning algorithms (such as the work of [3]). Other works prioritize the maintenance of good precision and accuracy and neglect the correlation between data such as the work of [14]. In addition, most of the work supports only monoculture. To summarize, at the current state, there is not a complete and accessible source of data which combines weather observations, soil data and estimated water quantity. Therefore, in the present work, we propose to complete the state of the art by offering a complete data source made up of the characteristics mentioned.

This source will be the building block for the water quantity prediction approach for our irrigation system which will, in turn, support a variety of crops. Through our approach, we aim to ensure the optimization of irrigation costs and resources.

5.4 CropWaterNeed Approach

In this section, we will present a general overview of the CropWaterNeed approach that we have established through some modifications of the classic machine learning process.

5.4.1 Data Collection

As part of the PRECIMED project, the major objective is to propose an approach based on machine learning for predicting the amount of water to irrigate. This approach should

operate on massive data, drawn from various sources and IoT systems. These systems are in the process of being purchased and are expected to provide various types of data relating to the farm and its environment (soil, weather, etc.). Due to the COVID19 crisis and the particular epidemiological circumstances, we could not move forward in setting up these systems, so we were forced to examine data sources for related work. However, as we have already pointed out the related work section, we notice the absence of a data source available for meteorological, soil and irrigation data. Therefore, we investigate a deep analysis of the literature review to identify how researchers have created their own database. Similar to these works, we prepared a set of data containing information on suitable characteristics for forecasting irrigation water demand. We built our database by combining on the one hand, CLIMWAT [16] which provides monthly climatic data on 144 countries and on the other hand, the values related to irrigation calculated by CROPWAT. Our choice is focused on CLIMWAT since it is offered by the United Nations Food Organization and agriculture (FAO) and all station information is taken from the FAO Agromet Group database. This source provides data weather, which include, temperature, humidity, precipitation, as well as the characteristics of the soil. When creating our database, irrigation needs calculations using data on climate, crops and soils as well as data on irrigation and the rain. The required climatic data are the reference evapotranspiration (monthly/10-year) and precipitation (monthly/10-year/daily). Evapotranspiration reference can be calculated from actual temperature data, humidity, sunshine/radiation and wind speed, according to the Penman method- FAO Monteith [17].

5.4.1.1 Collection of Meteorological Observations

During this step, we relied on CLIMWAT that provides average long-term monthly values of seven climatic parameters, namely:

- Average daily maximum temperature in °C
- Average daily minimum temperature in °C
- Average relative humidity in %
- Average wind speed in km/day
- Average number of hours of sunshine per day
- Average solar radiation in MJ/m²/day
- Monthly precipitation in mm/month
- Monthly effective rainfall in mm/month
- Reference evapotranspiration calculated by the Penman-Monteith method in mm/day.

Data can be extracted from single or multiple stations in a suitable format for their use in CROPWAT. Indeed, we plan to use meteorological observations data in combination with data from the CROPWAT program [18].

5.4.1.2 Release of Water Requirements

Water requirements for crops are essential data for irrigation prediction. To collect this kind of data, we relied on CROPWAT 8.0 for Windows. The latter is a program allowing to calculate crop water requirements and irrigation requirements from data on soil, climate and crops. It was developed by the Land and Water Development Division of FAO. In addition, the program allows you to build irrigation calendars for different management conditions and to calculate the supply of program water for different types of crops. CROPWAT can also be used to evaluate farmers' irrigation practices and to estimate performance crops under rainfed and irrigated cultivation conditions.

5.4.1.3 Data Aggregation

After collecting the meteorological observations and identifying the need for water, the goal is to bring all the data together in an Excel file. This file is then transformed into CSV format in order to obtain a database on which we can then apply machine learning algorithms.

5.4.2 Data Preparation

Following the construction of the database, preprocessing of the data set is necessary in order to make them useful for forecasting water demand and not interrupt the regression process. Data preparation includes analysis techniques of raw data in order to obtain quality data, in particular, collection, integration, transformation, cleaning, reduction and discretization of data [19]. As part of this research work, we have made changes on the characteristics by breaking down the "Date" characteristic into two characteristics: "Day" and "Month" to avoid the representation of dates in strings. Indeed, several machine learning algorithms, such as regression algorithms, require their inputs to be digital. Therefore, we replaced the string to numeric values. The data before and after the preparation is available on [20].

5.4.3 Feature Engineering

Feature engineering is considered to be one of the fundamental bases of machine learning which has a considerable impact on the performance of the model. As a general rule, the extraction characteristics consists of transforming the data into a vector of characteristics. Table 5.1 presents an extract of the used characteristics as well as their calculation methods. We choose the characteristics that have a significant influence on the use of water by crops and whose data are available throughout the seasons of culture. Our dataset contains historical data composed of characteristics on various weather observations such

Table 5.1 Extract of the used features

Category	Attribute	Description	Formula
Meteorological data	Rain	The amount of rain on the corresponding day	–
Soil data	Depl	The cumulative depth of evapotranspiration (depletion) from the rootzone (mm)	–
Irrigation data	Eta	Real evapotranspiration of crops	$Eta = Etc/Ks$ where Etc = evapotranspiration cultures under standard conditions $Ks = \text{water StressCoefficient}$
	Loss	Irrigation water that is not accumulated in the soil i.e. surface runoff or percolation in mm.	–
	Flow	Water flow rate in l/s/ha	–

as maximum temperature and minimum (Max-Temp & Min-Temp), wind speed, humidity, precipitation and solar radiation, combined with the type of soil, the type of crop and the use of water by crops. In fact, we define three categories of characteristics: meteorological characteristics, soil characteristics and characteristics related to irrigation. The first and second category are used by CROPWAT to predict the amount of water to irrigate. The third one contains data relating to our context which is irrigation. We consider our data as a two-dimensional array where the columns are characteristics (digital) and lines are the records.

5.4.4 Model Selection, Training and Evaluation

We have used learning algorithms known by their robustness [21] namely Random Forest, Decision Tree, Gradient Boosting Regressor and XGBoost Regressor to constitute the model for predicting the quantity of water to irrigate. We trained the generated models by each machine learning algorithm using a set of training data made up of examples used during the learning process. In addition, we also ensured the parameters es adjustment (e.g. the weights) of a regressor using this dataset. Technically, all machine learning algorithms were implemented using Python 3.7.0. In particular, we used the Scikit Learn library [22] to create and adapt our models under Google Colaboratory (also known under the name Colab) [23]. Colab is a cloud service based on Jupyter notebooks which provides a fully configured environment for machine learning and free access to a robust GPU. Subsequently, we evaluated the performance of each algorithm using the Mean Squared Error (MSE) [24], the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) [25].

In the next section, we present the experimentation of our CropWaterNeed approach using measurements of error rates generated by the prediction.

5.5 Experimentation

In this section, we analyse the performance of the learning algorithms. To do so, a series of experiments is carried out

Table 5.2 Performance results of the different learning algorithms

Model	MSE	RMSE	MAE
XGBRegressor	0.001	0.066	0.004
Random Forest	4.057	0.499	0.249
Decision Tree	1.515	0.015	0
Gradient Boost Regression	0.007	0.081	0.005

to build the models of prediction the amount of water to irrigate using the four automatic learning algorithms: Random Forest, Decision Tree, XGBRegressor and Gradient Boost Regression. The four prediction models are trained and tested using training and test sets extracted from the dataset that we built. Table 5.2 shows the performance of different prediction algorithms in terms of MSE, RMSE and MAE. It is obvious from the results that the XGBRegressor model offers the best performance when compared to the other three machine learning algorithms. This shows that the model generated by the XGBRegressor algorithm is robust and powerful in its capacity in prediction of the amount of water to irrigate. So our work is the first one that employed the XGBRegressor in the context of predicting the quantity of water to irrigate for smart agriculture.

5.6 Conclusion

Today, with the emergence of artificial intelligence techniques, especially machine learning and the popularization of connected objects, agriculture has passed towards an era of intelligence. As a result, the concept of smart agriculture was born. This concept has succeeded in overcoming the limitations of conventional agriculture (intensive use of labour and resources).

In order to take advantage of artificial intelligence in the agricultural field, we conducted a review of work related to intelligent agriculture. Further to this study, we noted that data availability remains a major problem. Thus, most of the work only supports monoculture. To meet the limitations of the existing work, we have proposed in this work the CropWaterNeed approach. To do this, we have created a database combining weather observations, soil characteris-

tics and irrigation data for a variety of crops. Then, we extracted from this database two other databases containing meteorological observations and irrigation data respectively. Moreover, we built the model for predicting the amount of water to be irrigated by applying machine learning algorithms based on the selected characteristics. The evaluation of the performance of the applied algorithms has shown that the best algorithm to use in our context is XGBRegressor. To the best of our knowledge, our work is the first to use the XGBRegressor in the context of quantity prediction of water to irrigate.

At the end of this work, several research perspectives could be considered. We aim to apply the same process, in order to test the performance of other regression algorithms in this case algorithms support vector regression (SVR) and linear regression. In addition, we would like to study the impact of using deep learning techniques on the data collected for the results with those of machine learning.

Acknowledgments The authors gratefully acknowledge the General Secretariat for Research and Technology of the Ministry of Development and Investments of Tunisia under the PRIMA Programme. PRIMA is an Art.185 initiative supported and co-funded under Horizon 2020, the European Union's Programme for Research and Innovation. (project application number: 155331/I4/19.09.18).

References

1. K.G. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, Machine learning in agriculture: A review. *Sensors (Basel)* **18**(8), 2674 (2018). <https://doi.org/10.3390/s18082674>
2. L. Shi, Q. Duan, X. Ma, M. Weng, The research of support vector machine in agricultural data classification, in *Computer and Computing Technologies in Agriculture V. CCTA 2011. IFIP Advances in Information and Communication Technology*, ed. by D. Li, Y. Chen, vol. 370, (Springer, Berlin, Heidelberg, 2011), pp. 265–269
3. A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, A. Sharma, IoT and machine learning approaches for automation of farm irrigation system. *Procedia Comput. Sci.* **167**, 1250–1257 (2020). <https://doi.org/10.1016/j.procs.2020.03.440>
4. O. Adeyemi, I. Grove, S. Peets, Y. Domun, T. Norton, Dynamic neural network modelling of soil moisture content for predictive irrigation scheduling. *Sensors (Basel)* **18**(10), 3408 (2018). <https://doi.org/10.3390/s18103408>
5. A.P. Patil, P.C. Deka, An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Comput. Electron. Agric.* **121**, 385–392 (2016). <https://doi.org/10.1016/j.compag.2016.01.016>
6. T.M. Mitchell, *Machine Learning*, 1st edn. (McGraw-Hill, Inc., New York, 1997)
7. T. Cover, Estimation by the nearest neighbor rule. *IEEE Trans. Inf. Theory* **14**(1), 50–55 (1968). <https://doi.org/10.1109/TIT.1968.1054098>
8. M. Segal, Machine learning benchmarks and random forest regression. Technical report, Center for Bioinformatics and Molecular Biostatistics. University of California, San Francisco, 2003.
9. A. Prasad, L. Iverson, A. Liaw, Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* **9**, 181–199 (2006). <https://doi.org/10.1007/s10021-005-0054-1>
10. M. Schonlau, Boosted regression (boosting): An introductory tutorial and a Stata plugin. *Stata J.* **5**, 330–354 (2005). <https://doi.org/10.1177/1536867X0500500304>
11. N. Duffy, D. Helmbold, Boosting methods for regression. *Mach. Learn.* **47**, 153–200 (2002). <https://doi.org/10.1023/A:1013685603443>
12. J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **28**, 337–407 (2000). <https://doi.org/10.1214/aos/1016218223>
13. J. Friedman, Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001). <https://doi.org/10.1214/aos/1013203451>
14. C. Salim, N. Mitton, Machine learning based data reduction in WSN for smart agriculture. *Adv. Intell. Syst. Comput.* **1151**, 127–138 (2020)
15. A.K. Moussa, M. Al-Janabi, M. Al Salam, Fuzzy based decision support model for irrigation system management. *Int. J. Comput. Appl.* **104**, 14–20 (2014). <https://doi.org/10.5120/18230-9177>
16. M. Smith, Food and Agriculture Organization of the United Nations, *CLIMWAT for CROPWAT: A Climatic Database for Irrigation Planning and Management* (FAO, Rome, 1993)
17. R. Allen, L. Pereira, D. Raes, M. Smith, FAO irrigation and drainage paper no. 56. Rome Food Agric. Organ. United Nations **56**, 26–40 (1998)
18. D. Clarke, M. Smith, K. El-Askari, *CropWat for Windows: User Guide* (IHE, Delft, 2000)
19. S. Zhang, C. Zhang, Q. Yang, Data preparation for data mining. *Appl. Artif. Intell.* **17**, 375–381 (2003). <https://doi.org/10.1080/713827180>
20. M. Fredj, R. Grati, K. Boukadi, CropWaterNeed dataset (2020) [Online], <https://github.com/malekfredj/cropwaterneed>. Accessed Nov 2020.
21. V. Rodriguez-Galiano, M. Sánchez Castillo, M. Chica-Olmo, M. Chica Rivas, Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **71**, 804–818 (2015). <https://doi.org/10.1016/j.oregeorev.2015.01.001>
22. F. Pedregosa et al., Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2012)
23. E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (Apress, Berkeley, 2019)
24. Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **26**(1), 98–117 (2009). <https://doi.org/10.1109/MSP.2008.930649>
25. T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **7**, 1247–1250 (2014). <https://doi.org/10.5194/gmd-7-1247-2014>

Clara Burbano, David Reveló, Julio Mejía, and Daniel Soto

Abstract

In a broad sense, Machine Learning (ML) is the performance optimization in a certain task through computational means, following a certain criterion and using referential data and/or past results from previous iterations. ML is a subset of Artificial Intelligence (AI) and has attracted a substantial amount of research during the last decades. This blooming subject led to the statement of different definitions for classifications, criteria, algorithms and so on. This paper summarizes these different definitions and proposes a homologation between them, providing an unified vision for each definition.

Keywords

Machine learning · Artificial intelligence · Criteria classification · Supervised learning · Unsupervised learning · Reinforced learning

6.1 Introduction

6.1.1 On the Growing Popularity of Artificial Intelligence and Machine Learning

During the last years, Artificial Intelligence (AI) related projects, as well as Machine Learning (ML) related projects grew in numbers, both at scientific and at industry level. While any developing research field sparks scientific interest on its own because of much knowledge is yet to be explored – with a quick Internet search, you can even find websites proposing research ideas [1–4], AI and ML proved to also be very attractive to different markets because of how much value companies get (and are forecasted to keep getting) thanks to AI and ML related products, proven by the growing marked size for these kind of technologies, predicted to be as big as US\$160 Bn by 2026 [5], as well as by how the stock markets for these fields have been positively behaving (and predicted to keep doing so) during recent years [6].

Solutions like customer behavior analysis, chatbots, business forecasting tools, behavior-based cybersecurity systems, to name a few, turned out to be highly profitable business [7, 8], which in turn made investment on these areas to grow as now is possible to even find crowdfunding websites for these kind of technologies [9].

This strong trend led to the involvement of many of the world’s largest companies, to active development AI and ML technologies. Interestingly enough, many of the companies leading in AI and ML markets, are also active players in the Cloud Computing industry, often developing hybrid solutions using AI and ML technologies on cloud services, promoting even more investing in these fields [10].

Current market analysis forecast that AI and ML could lead to an weighted average of 1.7% across 16 different

C. Burbano (✉)
Institución Universitaria Antonio José Camacho, Cali, Colombia
e-mail: clurbano@admon.uniajc.edu.co

D. Reveló · J. Mejía
Grupo de Investigación en Sistemas Inteligentes (GISI), Corporación
Universitaria Comfacauc, Popayán, Colombia
e-mail: drevelo@unicomfauca.edu.co;
jmejia@unicomfauca.edu.co

D. Soto
Centro de Investigación de la Universidad Mayor CICS, Universidad
Mayor, Santiago, Chile
e-mail: daniel.sotoca@mayor.cl

industries as well as to increase the economic output of those industries up to US\$4 trillion by 2035. What's more, when analyzed at country-level, it is forecasted that AI and ML could double the economic growth rates, among 12 countries sampled [11, 12]. It is also noteworthy that AI and ML are already creating new jobs, with industries requiring workers such as AI engineer, machine learning scientist, AI developer, among others [13].

The AI and ML have a well justified popularity both in the scientific and industrial communities. That being said, it is important to clarify and understand the difference between both, which is explained in the next Sect. 6.1.2. Later in Sect. 6.2 this paper dives into the different classification criteria used for ML algorithms, and then the Unified Vision Proposal is provided and explained in Sect. 6.3, followed up by conclusions and future work suggestions.

6.1.2 Defining Artificial Intelligence and Machine Learning

While the previous section mentioned AI and ML together, these are two different – but closely related – terms.

Although it is possible to – extremely – simplify this by stating that ML is a subset inside the AI field, brief but proper definitions are provided and referenced as follows:

Artificial Intelligence (AI). AI represents a set of complex edge technologies capable of interacting with its environment by means of simulating human intelligence [14] and is considered the core of the so-called “Fourth Industrial Revolution” [15].

Machine Learning (ML). Is the performance optimization in a certain task through computational means, following a certain criterion and using referential data and/or past results from previous iterations [16]. ML comes from the need to tackle problems beyond the reach of traditional, hardcoded IA solutions, being technically a specialized subset of AI, focused on real world knowledge applied to machines capable of making “subjective” decisions [17].

In a broad sense, the basic machine learning process involves building a ML based by “training” the machine using referential data [18].

6.2 Classifications and Selection Criteria in Machine Learning

6.2.1 Machine Learning Algorithm Classifications

Many authors concur in classifying ML algorithms based on a cognitive criterion, meaning that each ML algorithm

belongs to a certain group depending on how it “learns”. This approach identifies three main categories: Supervised learning, unsupervised learning and reinforced learning [18, 19], although some authors reduce these categories to the first two [17, 20].

Supervised Learning. As the name suggests, ML algorithms are “guided”. This guidance takes the form of referential data, usually called “target” data, so the algorithm knows that it must identify that kind of data. The algorithm then is trained used that target data, so when it is ready, it can identify whenever it is shown the target data or something else. This kind of algorithm is usually seen in tasks which require identifying what kind of input data (an image, for example) is being presented to the algorithm.

Unsupervised Learning. Unlike the previous ML type, those algorithms belonging to the unsupervised learning classification do not have the help of target data, so they rely on identifying patterns and structures on the input data they have to work with. In Example, an unsupervised ML algorithm will be given a set of pictures of pencils, apples and cars, so after iterating over that info, it will eventually be able to separate the pencils from the apples and the cars.

Reinforced Learning. These kinds of ML algorithms work similarly to its counterpart in psychology. The result of a task will be awarded or penalized depending on whether the answer is right or wrong, so the algorithm will learn from its previous experiences to answer right. Instead of using a target data set, it works with goals it aims to achieve. Some video games provide a very nice example of this kind of learning, when a character needs to go from point A to point B, while having many possible paths, by only one is the optimal one [21].

Other classification methods are based on the type of problem to be solved, or the type of data needed to be handled, or even in the type of statistical procedure required to achieve a solution. The strategies are closer to the decision criteria used to decide when to use a given algorithm, so we'll cover those in the next section.

6.2.2 Criteria for Choosing the Right ML Algorithm

While ML algorithms can be very flexible, some are more suited to certain scenarios than others.

Current literature states that the following variables are to be considered when deciding which algorithm is to be used:

- **Data size:** As some algorithms can have higher execution times, for very large datasets this can discard some options.

- Data quality: Algorithms relying heavily on the accuracy of the data presented to them (like in the case of supervised learning types), when the available data isn't reliable enough, it should be preferred to use algorithms which don't have this heavy dependency.
- Available time: Closely related to the data size variable, when confronted to a short deadline, some algorithms can be an actual obstacle to the research.
- Data type: Discrete and continuous data are to be approached differently, thus the algorithm must be chosen with this variable in mind as well.

6.3 Classification Filtering and Unified Vision Proposal

As previously stated, current literature provides a myriad of classification names to ML algorithms, often being redundant by giving a similar name to something already classified as a different name.

We have gathered all the definitions we found in the aforementioned literature, then filtered repeated results, and sorted them as a unified vision. Given the large number of concepts and relationships involved, the full scheme is presented in four parts, as shown in Figs. 6.1, 6.2, 6.3 and

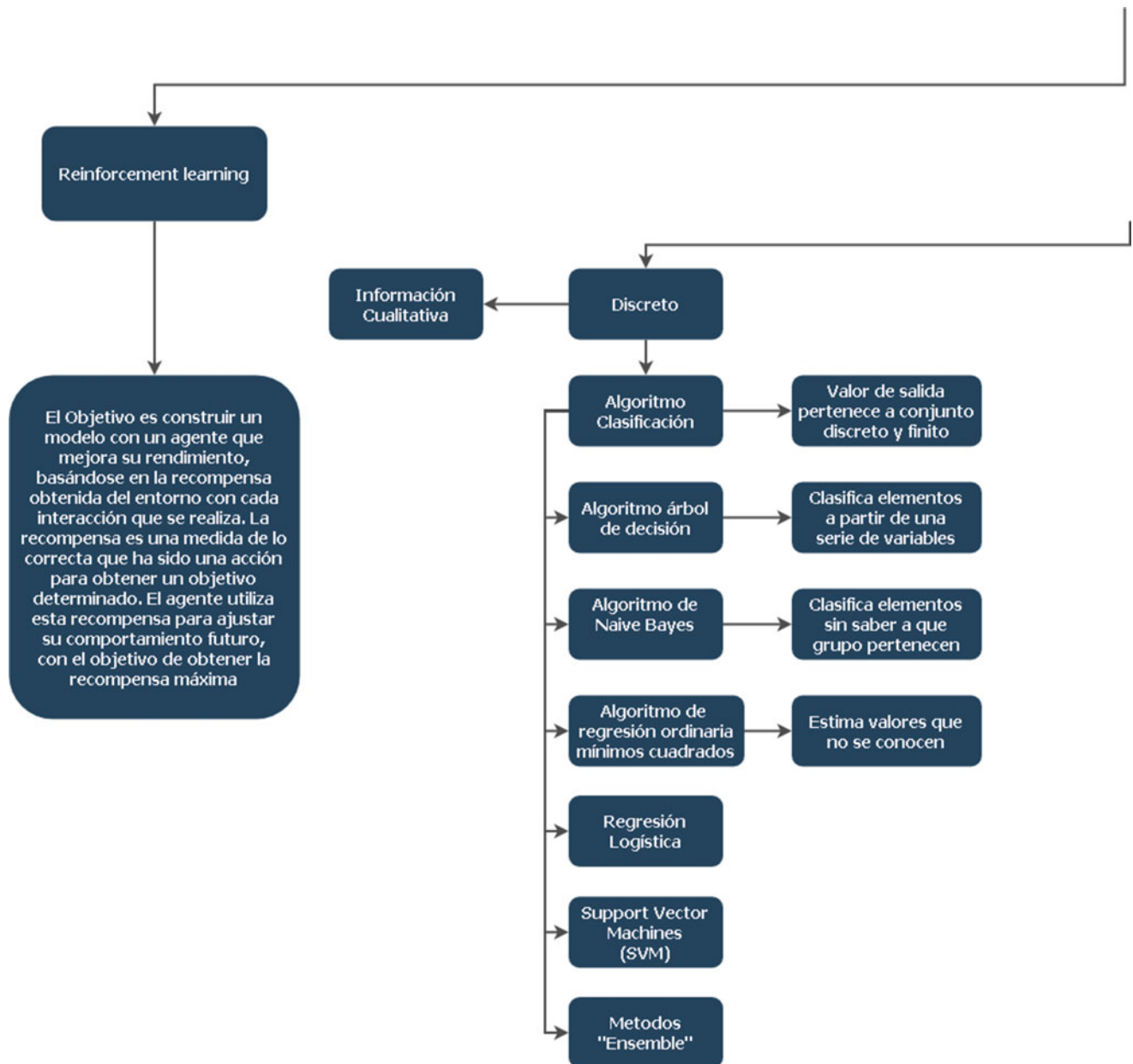


Fig. 6.1 Raw scheme, part 1. (Source: Prepared)

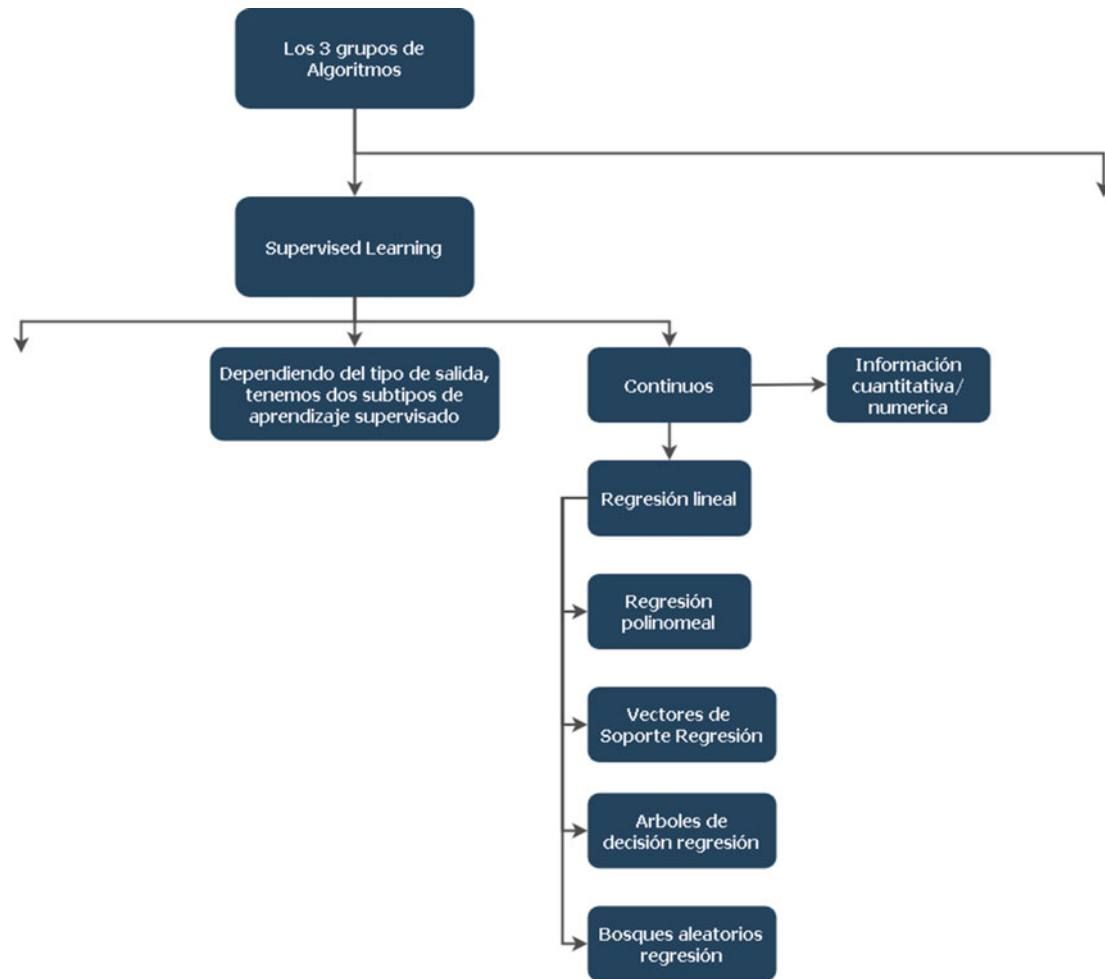


Fig. 6.2 Raw scheme, part 2. (Source: Prepared)

6.4. Then we sorted those relationships in a cleaner, shorter version, which is shown in Fig. 6.5.

6.4 Conclusions and Future Work

This research went through different views when it comes to how to classify ML algorithms, as well as views on which factors include to decide which algorithm is the best option for a given scenario. Then we found some common ground among various views, from which we graphically described how each view integrates into a greater scheme of things. It was possible to filter “repeated” views so as a result, we produced a refined version of this graphic perspective, in

the form of a conceptual map, which we propose as a tool to contribute to a better understanding of ML in a more structured way.

Nevertheless, by the time our research finished, new literature added even more views [22] up to 14 different ML algorithm types [23], so this proposal still has room for improvement. Future work should review those new classification proposals, in order to find a way to integrate them into this new, greater classification scheme.

Of course, as a relatively young and unexplored field, ML might lead to new algorithms and classifications currently unexplored, which may or may not integrate seamlessly into this scheme, thus our proposal might (or might not) require a deep reforming.

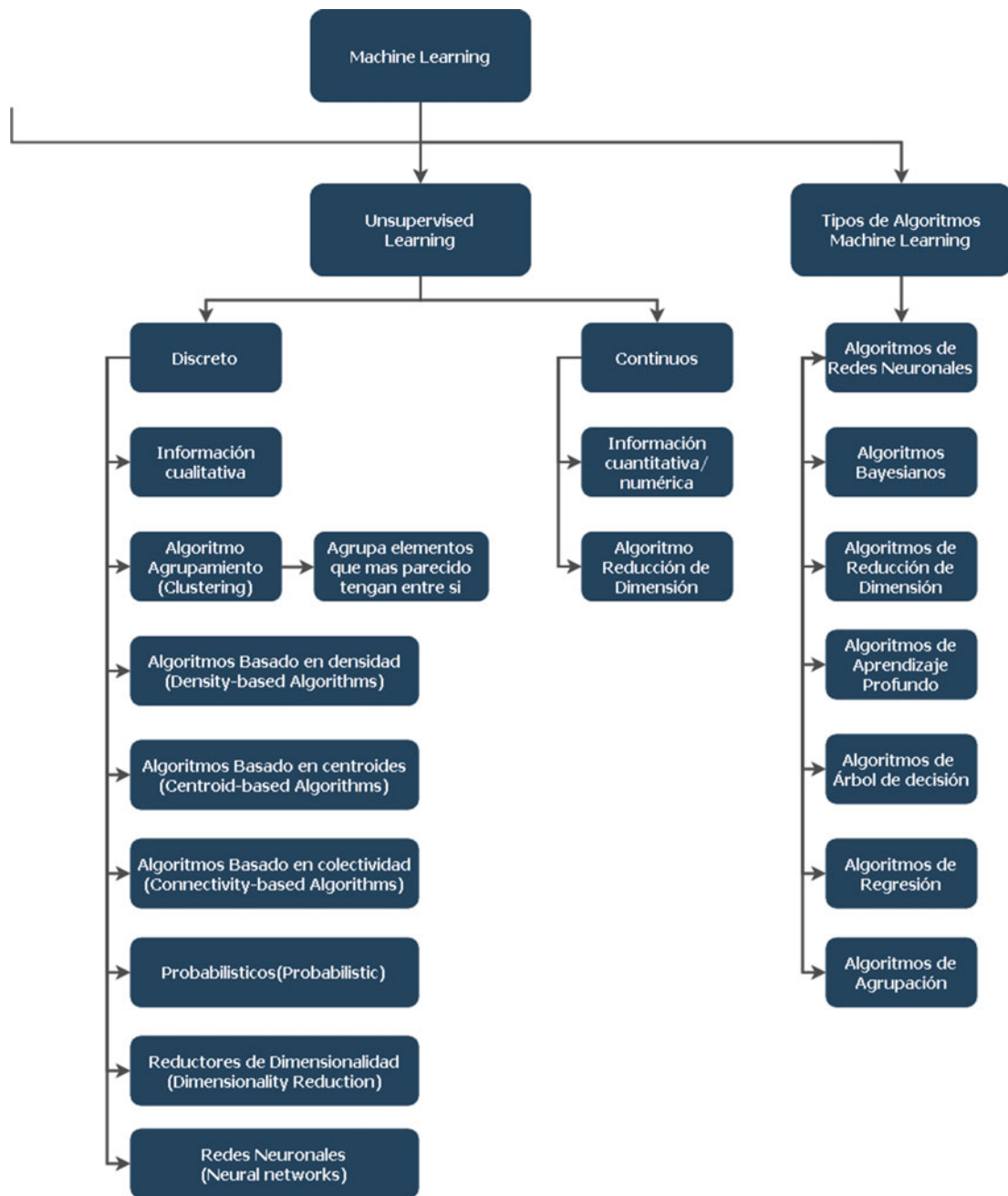


Fig. 6.3 Raw scheme, part 3. (Source: Prepared)

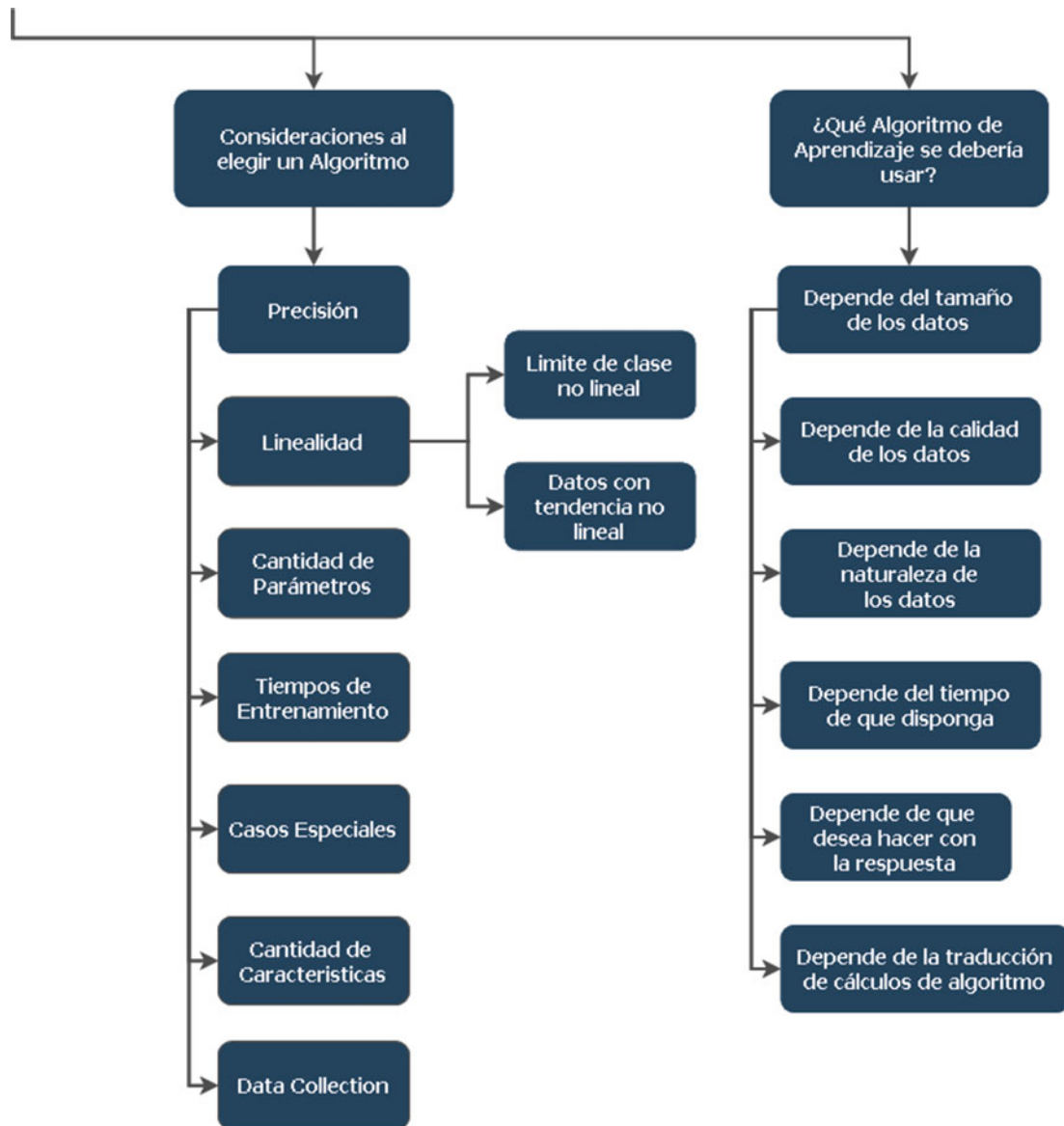


Fig. 6.4 Raw scheme, part 4. (Source: Prepared)

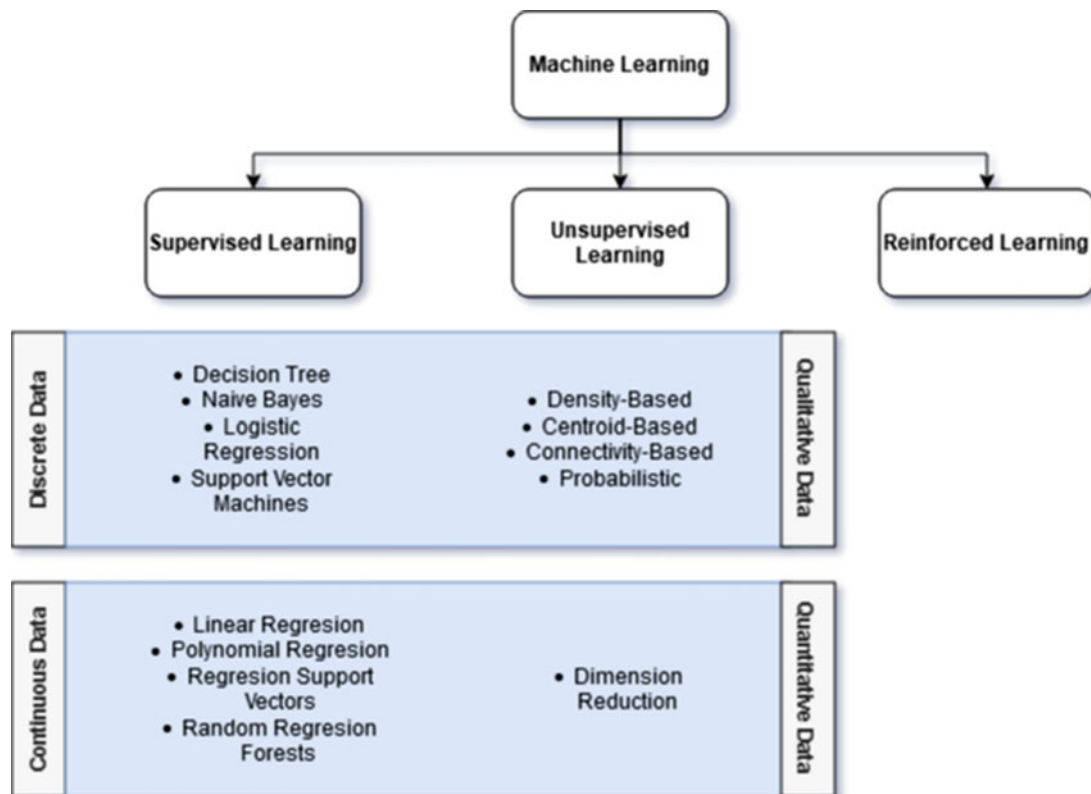


Fig. 6.5 Filtered scheme. (Source Prepared)

References

- Geeks for Geeks Website, <https://www.geeksforgeeks.org/8-best-topics-for-research-and-thesis-in-artificial-intelligence/>. Last accessed 2020/05/07
- 1 Red Drop Blog, <https://1reddrop.com/2020/05/14/15-artificial-intelligence-research-paper-topics-for-writing/>. Last accessed 2020/05/07
- Tech Sparks Blog, <https://www.techsparks.co.in/artificial-intelligence-as-an-m-tech-thesis-topic-for-cse/>. Last accessed 2020/05/07
- Data Flair Blog, <https://data-flair.training/blogs/machine-learning-project-ideas/>. Last accessed 2020/05/07
- Market Watch Press Release, <https://www.marketwatch.com/press-release/artificial-intelligence-ai-market-value-growing-at-a-us-160-bn-by-2026-2020-04-13>. Last accessed 2020/05/07
- The Motley Fool Website, <https://www.fool.com/investing/stock-market/market-sectors/information-technology/ai-stocks/>. Last accessed 2020/05/07
- Towards Data Science Website, <https://towardsdatascience.com/value-investing-with-machine-learning-e41867156108?gi=1840c5a0962c>. Last accessed 2020/05/07
- T.A. Borges, R.F. Neves, Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods. *Appl. Soft Comput.* **90**, 106187 (2020)
- C.C. Chen, C.H. Chen, T.Y. Liu, Investment performance of machine learning: Analysis of S&P 500 index. *Int. J. Econ. Financ. Issues* **10**(1), 59–66 (2020)
- Datamation Website, <https://www.datamation.com/artificial-intelligence/top-artificial-intelligence-companies.html>. Last accessed 2020/05/07
- Accenture Website: How AI boosts industry profits and innovation, https://www.accenture.com/_acnmedia/Accenture/next-gen-5/insight-ai-industry-growth/pdf/Accenture-AI-Industry-Growth-Full-Report.pdf?la=en. Last accessed 2020/05/07
- Accenture Website: Industry spotlights: How AI boosts industry profits and innovation, https://www.accenture.com/_acnmedia/Accenture/next-gen-5/insight-ai-industry-growth/pdf/Accenture-AI-Industry-Growth-Industry-Report.pdf?la=en. Last accessed 2020/05/07
- Datamation Website, <https://www.datamation.com/artificial-intelligence/artificial-intelligence-jobs.html>. Last accessed 2020/05/07
- E. Glikson, A.W. Woolley, Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **14**(2), 627–660 (2020)
- K. Schwab, *The Fourth Industrial Revolution* (Crown Business, New York, 2017)
- E. Alpaydin, *Introduction to Machine Learning* (MIT Press, 2020)
- I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016)
- Edureka! Blog, <https://www.edureka.co/blog/introduction-to-machine-learning/>. Last accessed 2020/05/07
- Medium Article: Different types of machine learning and their types, <https://medium.com/deep-math-machine-learning-ai/different-types-of-machine-learning-and-their-types-34760b9128a2>. Last accessed 2020/05/07

20. Hunter Heidenreich Blog: Machine learning for the average person: What are the types of machine learning?, http://hunterheidenreich.com/blog/breaking_down_ml_for_the_average_person/. Last accessed 2020/05/07
21. M. Kubat, B. Ivan, M. Ryszard, A review of machine learning methods, in *Machine Learning and Data Mining: Methods and Applications*, (Wiley, Chichester, 1998), pp. 3–69
22. A. Dey, Machine learning algorithms: A review. *Int. J. Comput. Sci. Inf. Technol.* **7**(3), 1174–1179 (2016)
23. Machine Learning Mastery Blog, <https://machinelearningmastery.com/types-of-learning-in-machine-learning>

Part II

Cybersecurity I

Classification and Update Proposal for Modern Computer Worms, Based on Obfuscation

Hernaldo Salazar and Cristian Barria

Abstract

Computer worms are a type of malware that have a complex technological structure and the ability to automatically create replicas of themselves without human interaction and distributing themselves to other computers connected to the network; they have a malicious code component that allows them to infect one computer and then use it to infect others. This cycle repeats itself, rapidly increasing the number of infected computers if action is not taken in time. Within this framework, the research is based on a systematic review of the methodology used to analyze scientific articles related to malware and specifically to computer worms. Through this review and the abstraction of important data, a synthesis of the results is made to support the research, resulting in a new proposal for the classification of computer worms according to their obfuscation capacity, dividing it into four levels: species, type, class and evasion. This classification allows a modern computer worm to be categorized in such a way that the main contribution is that it can serve as a model or as a complement to an Information Security Management System (ISMS), in the systems responsible for detecting and/or defending organizations against worms attacks.

Keywords

Classification · Cybersecurity · Detection · Malware · Program · Propagation · Self-replicating · Specie · System · Taxonomie

H. Salazar (✉) · C. Barria
Escuela de Ingeniería en Computación e Informática, Universidad Mayor, Providencia, Chile
e-mail: hernaldo.salazar@mayor.cl; cristian.barria@umayor.cl

7.1 Introduction

Computer worm is a self-replicating program (SRP), considered to be within the category of Malware [1], and therefore is a malicious code that has the main intention of causing damage to a computer system, using the network to make copies of itself, spreading to other hosts (computers connected to the network), executing this action without the user's participation [2, 3]. From a historical perspective the first computer worm was Creeper. Ravinder Nellutla who claims [4]:

In the early 1970s and written by Bob Thomas, it was an experimental program to demonstrate the power of programming. Most of the worms written at that time were the result of programmers' fascination with self-replicating programs. There was no malicious intent, and the worms did not hide. They were sent in clear. The Creeper worm was written to infect DEC PDP-10 (Digital Equipment Corporation Programmed Data Processor model 10) computers running the TENEX operating system. The program used ARPANET to spread from one node to another and display a message "I'm the creeper, catch me if you can!". A Reaper program was written to counteract Creeper.

Despite this, it is known that the idea of the SRP precedes Creeper by many years. In this regard, Thomas M. Chen states [5]:

However, the idea of self-replicating programs dates back to early 1949 when mathematician John Von Neumann imagined specialized computers or "self-replicating automatons" that could build copies of themselves and transmit their programming to their progeny.

Over the years, the SRP varied in structure, behavior, form of attack, among others, but always with capabilities to circumvent security functions. This has led to different

classifications depending on the computer worms studied. One classification that has emerged due to the evolution of viruses and worms is that proposed by Tomas Chen in a 2004 study [5]:

The history of the evolution of viruses and worms is classified into four “waves” ranging from the Shoch and Hupp worms in 1979 to the present (the term “generations” is less preferred because viruses and worms of one wave are not direct descendants of a previous wave). This historical perspective determines a series of observations about the current state of vulnerability and trends in possible worm attacks.

We classify the evolution of viruses and worms in the following waves:

- First wave, from 1979 to the early 1990s
- Second wave, from the early 1990s to 1998.
- Third wave, from 1999 to 2001.
- Fourth wave, from 2001 to date.

This classification serves as a reference and takes into account the changes that were to take place in the evolution of this type of malware. In this sense, they serve as a reference for the classification of modern worms.

7.2 The Problem

Given the lack of classifications of computer worms based on obfuscation patterns, the idea of finding differentiators between this type of malware arises. This should contribute, in the long run, to the integrity, availability and confidentiality of organizations. The use of an appropriate classification would make it possible to minimize the risk of possible attacks, obtain and index new information about their behavior, verify the level of the threat, and detect them in time [6].

Since each type of computer worm analyzed by researchers has different characteristics and different types of classification are carried out for them, we have a case, of a classification of computer worms raised by Madihah Mohd Saudi that states [7]:

After the analysis carried out in the laboratory, it leads the researchers to produce a new classification for the EDOWA system. A worm classification proposal is made. This classification is based on several factors: infection, activation, payload, operational algorithms and propagation.

This classification proposal is produced based on the research and tests that have been carried out in the laboratory. The classifications are divided into five main categories: Infection, Activation, Payload, Operational Algorithms and Propagation. The Efficient Worm Attack Detection System (EDOWA) is produced according to this classification.

It is clear that this research was based on the attack of a computer worm and not on a characteristic of the worm, so that overall margins of behavior are obtained as a result. Moreover, if we work on the prediction of the infectious nodes it would be possible to reconstruct the scene of the computer worm attack and identify the origins of the spread [8].

7.3 Related Work

This section presents previous work associated with malware, obfuscation and computer worms, related to this work. Firstly, a focus on worm detection systems. Next, research is presented on a proposed classification of malware based on obfuscation and finally a method for investigating the spread of worms in the network.

Worm detection systems. Unlike the traditional worm detection system (WDS) [9], which mainly uses a signature and behavior-based approach, this approach is based on detection according to the damage caused. In order to find common ground for this type of threat, there must also be a systematic and methodological analysis process [10], to acquire knowledge about a particular malware. Therefore, the code analysis phases (both static and behavioral) can be focused on the victim system, simplifying the number of tools to be installed on the rest of the systems involved in the laboratory. This allows the analyst to focus on the objective of the monitoring system and the system services as a key part of a realistic scenario to force malware to show its full behavior and thus have more accuracy in detecting any malware on a host [11].

Malware classification based on obfuscation. The literature offers a classification of malware from multiple perspectives, but there is no obfuscation classification that allows us to match a malware with the given capabilities, so it is imperative to structure it hierarchically so that the capabilities relevant to each are known [12]. Due to the rapid evolution of malware, it is not possible to maintain the same type of classification for long. They need to be classified by finding unstudied common patterns that allow a more immediate understanding of their functioning and characteristics. It can even be manual and automated to have a classification [13]. In the case of the automated method, time and cost resources are used more efficiently. All of this implies a challenge for those who develop malware as they have to deal with those who protect and those who evade data protection systems [14].

Spread of worms in the network. Identifying the origins of a worm’s propagation route is very useful for those in charge of detecting possible suspects, finding out what security weaknesses there are in the network and discovering the nodes through which the worm enters [15]. The main idea

is to identify the origins and reconstruct the propagation path of the worm attack using information gathered from the network. The main objective is to reduce the limitations of current methods, such as computational complexity and storage requirements. Four approaches are taken to identify the origin and reconstruct the propagation path. On the other hand, an investigation provides us with information about the problems of detecting computer worms in networks [16]. Many existing schemes use a single parameter for detection that leads to a poorer characterization of the threat model, therefore, a high rate of false positives and false negatives. In short, not all computer worms can be detected with a single method [17].

7.4 Methodology and Classification Proposal

The present research is based on a systematic review as a methodological strategy of analysis for data abstraction. The process begins with the systematic review, after which the databases that will be used for the review are chosen, which in this opportunity are seven; inclusion and exclusion criteria are formulated for the search chains, since we work with a range not greater than 5 years; after the initial reading, the inclusion and exclusion of titles and abstract of scientific articles are made; a review of the available ones is made and the information is abstracted with relevant data; later a critical analysis of the articles is made, and a study summary that gives way to a synthesis of results. This allows us to have a basis for identifying and analyzing relevant work.

The sample of this work is related to the scientific articles selected as a result of the search in the academic databases (Table 7.1), representing the first part of the systematic review. The table shows the distribution of the study sources where the criteria and the years for the searches carried out are observed. The initial number of scientific articles found is 235 using two strings for these searches (Table 7.2).

The procedure used was first to obtain a sample with the extraction by means of two search chains in seven databases, obtaining a total of 235 scientific articles, from which it can be inferred that they mostly belong to the Netherlands, United States, United Kingdom and Egypt (Fig. 7.1).

Afterwards, an initial reading of the titles and abstracts is done to find an inclusion and exclusion process, reaching a quantity of 40 articles available for the revision of the full text, which will be done for a further deepening of the results and the establishment of a support for the research.

From the full text review, 31 scientific articles are obtained that contribute to the research, duplicates are discarded and a quantity of 24 is obtained, associated with 12 countries (Fig. 7.2). Next, a categorization is made by year of publication and their respective academic databases (Table 7.3).

From the abstracted data of the scientific articles selected as a result of the systematic review (Table 7.3), criteria based on identifying, selecting and reviewing these in a period of time between the years 2015–2020, are used to seek and obtain the necessary information for the purpose of choosing those that present studies associated with the objective of this research.

The information obtained is organized into categories so that the main focus is on the structure, characteristics and patterns of behavior. A table is prepared to analyze the data obtained with the purpose of extracting conclusions that allow to obtain the expected result for the objective (Table 7.4).

7.5 Classification of Computer Worms

To understand the types of classifications, it is assumed that there are several approaches to determine a classification, as in the case of the TCP worms detection research [18] called NADTW (New Approach for Detecting TCP worms) that allows us to know more about this type of species, allowing us to know that a network worm uses as a first measure a network scan for the identification of vulnerable hosts and services.

Another possible classification is that of behavioral analysis in the initial stage [19], but with a focus on malware in general and not specifically on computer worms, which allows us to have a general understanding of the use of hybrid analysis to classify the binaries of six types of malware species. In this case the researchers used labels provided by Microsoft according to its Computer Antivirus Research Organisation (CARO) malware nomenclature.

There are three taxonomies that present us with classifications of worms with different criteria. These help to understand how researchers have addressed the characteristics of computer worms.

The first is a taxonomy of computer worms [20] based on development based on the discovery of targets, carriers, activation, payloads and attackers. The carrier, activation and payload are independent of each other and describe the worm itself. In this approach the focus is not only on the worms, but also on the attackers and their motivations, because worms are ultimately written by humans and sometimes the easiest way to defend against a worm is to eliminate the motivation to write it (Fig. 7.3).

A second taxonomy is that of Efficient Detection of Worm Attack (EDOWA) [7], which aims to divide the classification for computer worms into five main categories, these are Infections, Activation, Payload, Operational Algorithms and Propagation. To produce this new classification of worms, the authors conducted tests and research. They built a code analysis laboratory to test and analyze the worm, in a con-

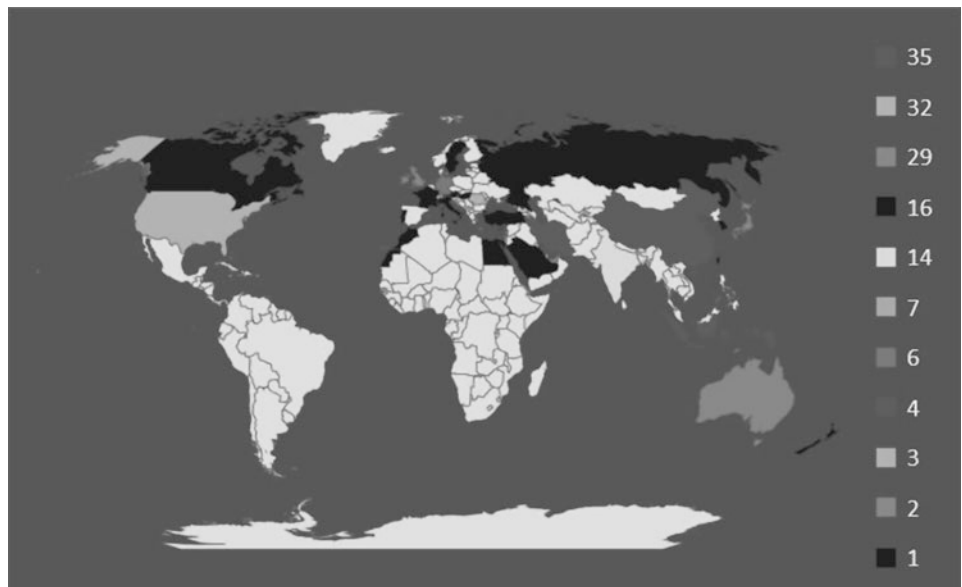
Table 7.1 Academic database search

Database	Search criteria	Years	Initial amount	Strings
Academic search ultimate	Worm, malware, obfuscat*, computer	2015–2020	15	2
Applied Science & Technology Source Ultimate			57	
Computer Source			6	
Computers & Applied Sciences Complete			26	
Engineering Source			20	
IEEE			60	
Science Direct			51	
Total			235	

Table 7.2 Strings used searches academic databases

Database	String 1	String 2
Academic search ultimate Applied Science & Technology Source Ultimate Computer Source Computers & Applied Sciences Complete Engineering Source	TI (malware OR worm (AND AB obfusca*)) AND AB (malware OR worm) AND TX (malware AND worm) AND TX (obfusca*)	TI (worm*) AND AB (worm* AND computer) AND TX (computer AND worm AND malware)
IEEE	(((((("Document Title":malware) OR "Document Title":worm) AND "Abstract":obfusca*) AND "Abstract":malware) OR "Abstract":worm) AND "Full Text & Metadata":malware) AND "Full Text Metadata":worm) AND "Full Text & Metadata":obfusca*)	(((((("Document Title":worm*) AND "Abstract":worm*) AND "Abstract":computer) AND "Full Text Only":computer) AND "Full Text & Metadata":worm)
Science Direct	Find articles: malware AND worm – Title, abstract, keywords: obfuscate OR obfuscating OR obfuscated OR obfuscation	Find articles: worm AND computer – Title, abstract, keywords: computer AND worm

Fig. 7.1 World map of congresses and papers found by searching academic databases



trolled environment. This laboratory is not connected to the real network. By using three machines and connecting them to the LAN via a hub, this resulted in a new classification (Figs. 7.4 and 7.5).

The third taxonomy is the so-called Computer Worm Classification [21], which attempts to demonstrate that the computer worm is not a simple malware. Therefore, research attempts to classify worms based on four main elements,

Fig. 7.2 Number of countries associated to the selected scientific articles to abstract the information

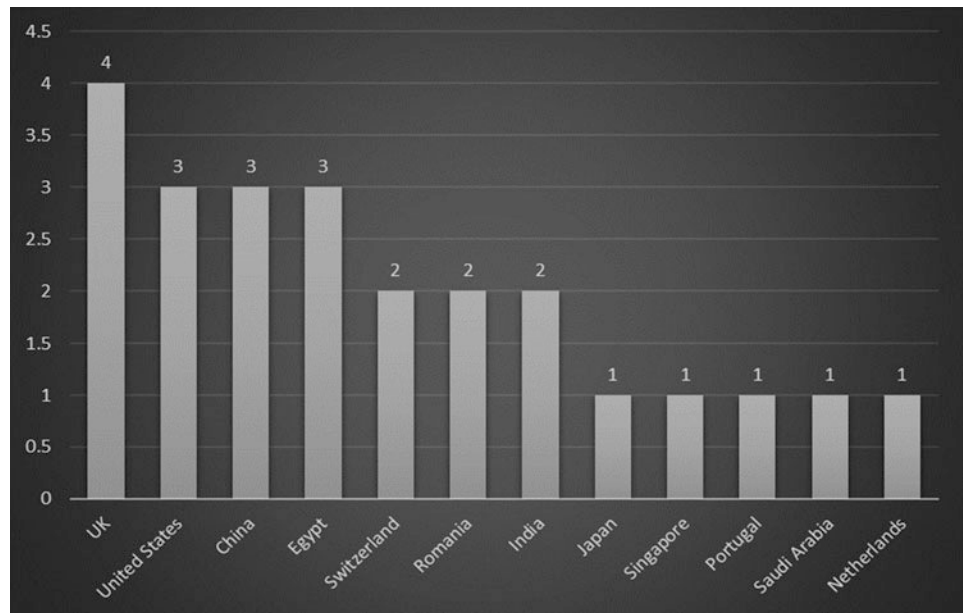


Table 7.3 Database distribution and years of publications

Database	2015	2016	2017	2018	2019	2020
Academic search ultimate	1		1			
Applied Science & Technology Source Ultimate	1	1	3	1	1	1
Computer Source			1			
Computers & Applied Sciences Complete	1		2			
Engineering Source	1		1			
IEEE	3	2	2	4	2	
Science Direct			1		1	
Papers per year	7	3	11	5	4	1
Duplicate papers	5	3	6	5	4	1
Total	24					

Table 7.4 Selected criteria distributed in categories

Category	<i>n</i>	%
Infection	22	13
Activation	10	6
Payload	12	7
Algorithms	20	12
Spread	16	10
Discovery	25	15
Exploitation	14	8
Hiding	12	7
Obfuscation	16	10
Behavior	20	12

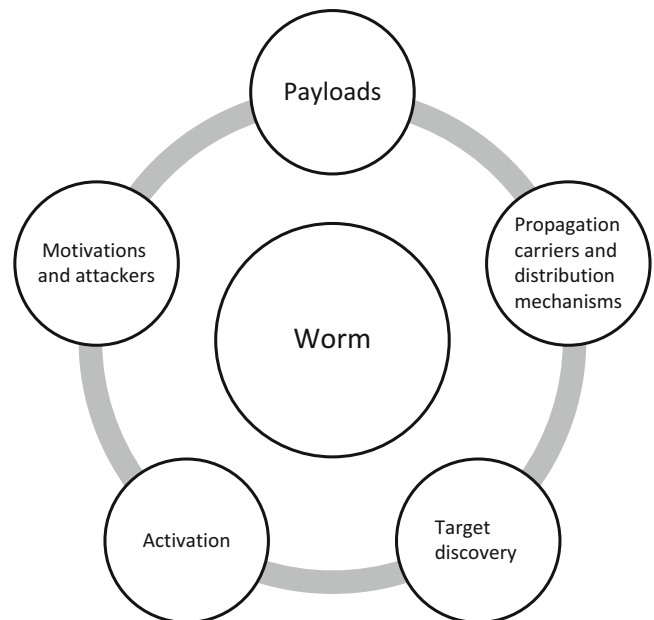


Fig. 7.3 Preliminary taxonomy classification of computer worms

classified as follows: worm structure, worm attack, worm defense and user defense. This gives a clearer understanding of the computer worm, how it acts and how to fight it (Fig. 7.6).

7.6 Proposed Classification Model

From the results obtained, as well as from the revision of the existing taxonomies, work begins to focus on the common

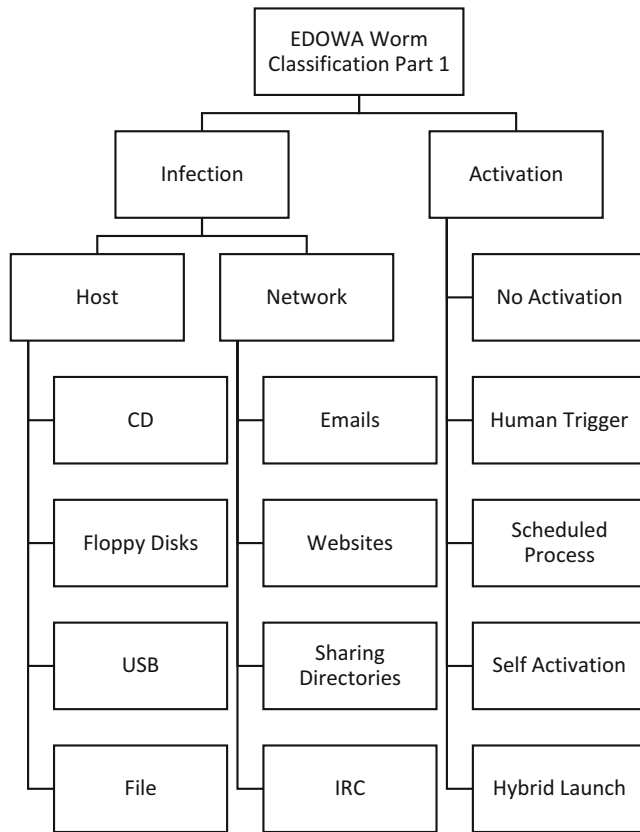


Fig. 7.4 Classification of computer worms EDOWA Part 1

criteria as a selection process to develop the classification model that allows us to categorize a computer worm by means of its obfuscation.

Today many people do not understand the differences between the different types of malware, and this is why the seriousness of this type of threat is increasing [22]. This lack of understanding and awareness of the dangers imposed on the systems we use must be addressed if we are to continue introducing these technologies into the workplace. For this type of thing, the solutions that help threat analysts must be easy to understand. We believe that this can minimize risks such as that represented by the Conficker computer worm in 2008 [23]. To date, researchers have analyzed its behavior, as it was a critical threat due to the hybrid nature of its spread, based on three strategies: local survey, neighborhood survey and global survey.

Possibly an insecure environment known as the Internet [24] are challenges for the protection of interconnected computer systems. It is a fact that malware has the potential to seriously affect any type of Internet user, from private companies to government institutions. Research has focused on analyzing the behavior of polymorphic malware.

The classification of computer worms is the first part of what could be a procedure for concealing this species, using obfuscation techniques [12]. To this end, work is being

done on the development of an inverted pyramid whose main category is the modern computer worm. The other levels are described below.

7.6.1 Class

The second level corresponds to the “classes”, whose main characteristic is propagation, and they correspond to (Table 7.5):

- Internet: Worms that spread through the Internet by exploiting security flaws [22].
- P2P: Worms that propagate in P2P networks by exploiting bugs [22].
- Email: Worms that spread via email by sending infected messages [22].

7.6.2 Type

The third level corresponds to the “types”. Their main characteristic is concealment, and they correspond to (Table 7.6):

- Polymorphic: Changes its binary code through encryption while keeping the original worm code intact [25, 26].
- Polymorphic exploitation: Worm infection consisting of exploit and Payload. You can change Payload dynamically through a polymorphic or metamorphic worm code, the same for the exploitation that mutates some bytes [21].
- Metamorphic: Capable of creating a new generation of worms in which the code is transformed or changed [27].
- Encryption: Capable of having its main body encrypted [12].

7.6.3 Evasion

The fourth level corresponds to “methods”. Their main characteristic is evasion, and they correspond to (Table 7.7):

- Based on signatures: It is a traditional approach, working with databases containing signatures of patterns or habits of the worms [9].
- Based on anomalies: It is an approach that builds the normal behavior model of the network or a program, alerting if this behavior changes [9].

Thus, there are four levels that allow structuring a classification for computer worms, in order to establish a taxonomy according to their species, class, type and evasion (Fig. 7.7).

A different taxonomy to those studied was one related to banking Trojans, which focused on threat intelligence based on the cyber kill chain, with details of their characteristics.

Fig. 7.5 Classification of computer worms EDOWA Part 2

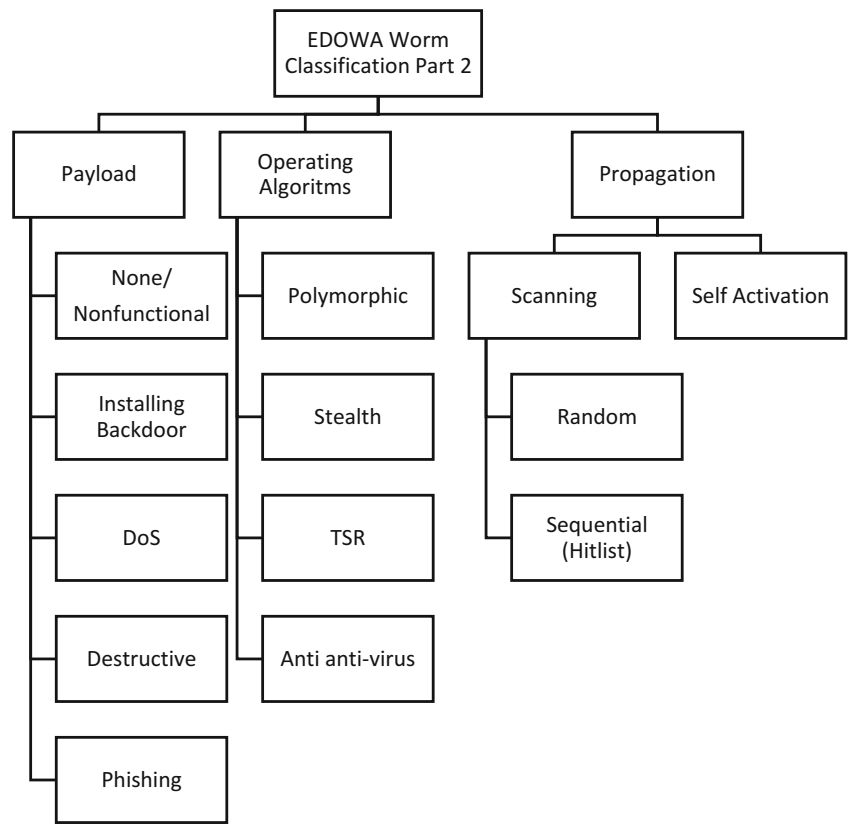


Fig. 7.6 Classification of computer worms into four categories

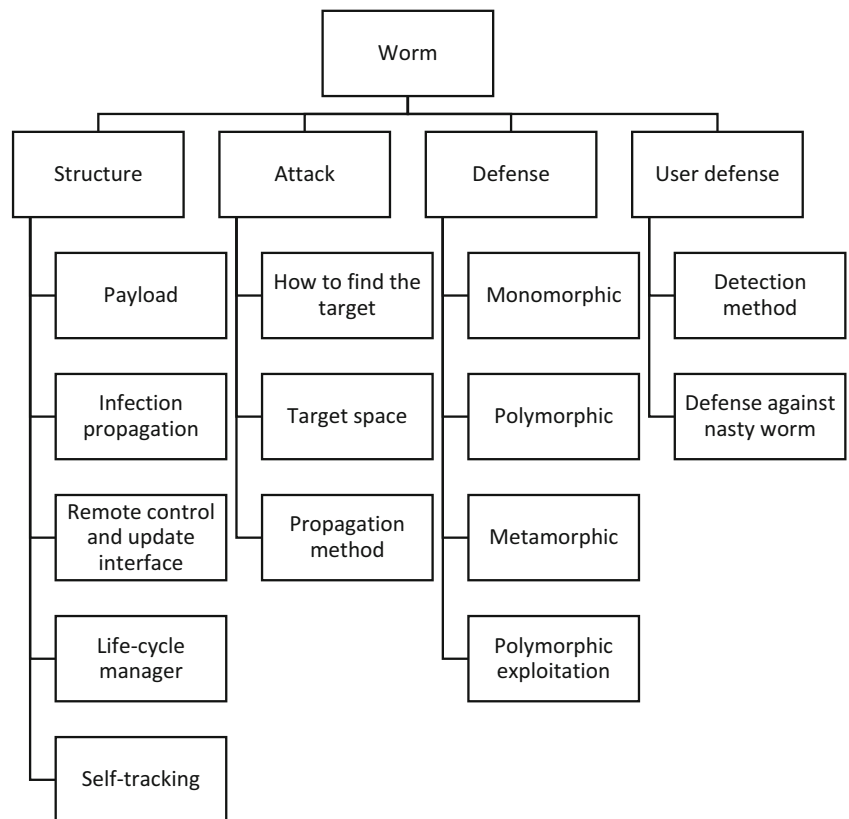


Table 7.5 Class summary

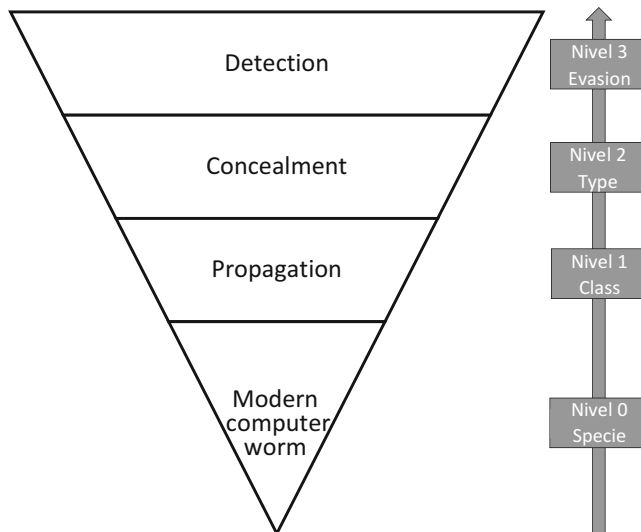
Propagation
Internet
P2P
Email

Table 7.6 Type summary

Concealment
Polymorphic
Polymorphic exploitation
Metamorphic
Encryption

Table 7.7 Evasion summary

Detection
Based on signatures
Based on anomalies

**Fig. 7.7** Proposal for classification of computer worms

The aim was to reduce the imprecision, subjectivity and uncertainty of knowledge in the decision-making process [28]. A taxonomy of the malware analysis tools associated with the security and privacy problems of IOs was also analyzed [29].

7.7 Conclusion and Future Work

The research reviewed above has helped to understand that classifications can be established, based on different perspectives so the classification proposed in this research has been based mainly on a synthesis of the evidence available in multiple scientific articles. This with the purpose of obtaining information and support that allows proposing a new

classification, with a hierarchical structure in the form of an inverted pyramid by levels, whose main characteristic is that it is based on patterns of obfuscation.

The various articles and books already in existence offer different points of view on the classification of malware and computer worms, due to their rapid evolution as malware developers focus mainly on implementing evasion techniques to avoid detection or their type of obfuscation that does not allow for observation of their code in plain text.

This paper proposes a classification that allows a computer worm to be catalogued in such a way that its main characteristic is when it uses a cipher method, obfuscating its code, thus obtaining an updated malware that falls into the category of modern computer worm and leaving out those that do not use this type of method, making them obsolete.

Future work could focus on classification in the field of threat intelligence, allowing us to work not only reactively to new threats but also proactively to new samples. Likewise, it is necessary to build a software or tool that allows us to obtain this result in a more automated way by means of a computer worm sample, with a view to forensic analysis, incident management or the search for threats associated with this species with the main characteristic “obfuscation”.

References

1. A. Tajoddin, S. Jalili, HM3alD: Polymorphic malware detection using program behavior-aware hidden Markov model. *Appl. Sci.* **8**(7), 1044 (2018)
2. V.S. Koganti, L.K. Galla, N. Nuthalapati, Internet worms and its detection, in *International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCI-CCT)*, 2016
3. L. Xue, Z. Hu, Research of worm intrusion detection algorithm based on statistical classification technology, in *8th International Symposium on Computational Intelligence and Design (ISCID)*, 2015
4. R. Nellutla, V.P. Goranthalala, F.A. Parvez, Classification of different computer worms with dynamic detection using victim number based algorithm. *Int. J. Eng. Res. Appl.* 803–812 (2013)
5. R. Chen, The evolution of viruses and worms, in *Statistical Methods in Computer Security*, (CRC Press, New York, 2004)
6. M. Khan, I.R. Khan, Malware detection and analysis. *Int. J. Adv. Res. Comput. Sci.* **8**(5), 1147–1149 (2017)
7. M. Saudi, E. Tamil, S. Nor, M. Idris, K. Seman, EDOWA worm classification, in *Lecture Notes in Engineering and Computer Science*, 2008
8. T. Tafazzoli, B. Sadeghiyan, A stochastic model for the size of worm origin. *J. Comput.* **9**(10), 1103–1118 (2016)
9. Y. Al-Saawy, A. Cau, F. Siewe, A novel approach to worm detection systems, in *2015 Science and Information Conference (SAI)*, 2015
10. J. Bermejo, C. Aramburu, J.-R. Higuera, M. Urban, J.A. Montalvo, Systematic approach to malware analysis (SAMA). *Appl. Sci.* **10**(4), 1360 (2020)
11. B.M. Khammas, S. Hasan, R.A. Ahmed, J.S. Bassi, I. Ismail, Accuracy improved malware detection method using Snort sub-signatures and machine learning techniques, in *10th Computer Science and Electronic Engineering (CEECE)*, 2018

12. C. Barría, D. Cordero, C. Cubillos, M. Palma, Proposed classification of malware, based on obfuscation, in *6th International Conference on Computers Communications and Control (ICCCC)*, 2016
13. C. Barría, D. Cordero, C. Cubillos, M. Palma, D. Cabrera-Paniagua, Obfuscation-based malware update, a comparison of manual and automated methods. *Int. J. Comput. Commun. Control* **12**(4), 461–474 (2017)
14. I. Shiel, S. O’Shaughnessy, Improving file-level fuzzy hashes for malware variant classification. *Digit. Investig.* **28**, S88–S94 (2019)
15. T. Tafazzoli, B. Sadeghiyan, A four-step method for investigating network worm propagation, in *7th International Symposium on Digital Forensics and Security (ISDFS)*, 2019
16. N. Ochieng, W. Mwangi, I. Ateya, Optimizing computer worm detection using ensembles. *Secur. Commun. Netw.* **2019**, 4656480 (2019)
17. D. Jain, S. Khemani, G. Prasad, Identification of distributed malware, in *IEEE 3rd International Conference on Communication and Information Systems (ICCIS)*, 2018
18. M. Anbar, R. Abdullah, A. Munther, M. Al-Betar, R. Alnakhalny, NADTW: New approach for detecting TCP worm. *Neural Comput. Appl.* **28**, 525–538 (2017)
19. N. Kumar, S. Mukhopadhyay, M. Gupta, A. Handa, S.K. Shukla, Malware classification using early stage behavioral analysis, in *14th Asia Joint Conference on Information Security (Asia JCIS)*, 2019
20. N. Weaver, V. Paxson, S. Staniford, R. Cunningham, A taxonomy of computer worms, in *WORM’03 – Proceedings of the ACM Workshop on Rapid Malcode*, 2003
21. A. Pratama, F.A. Rafrastara, Computer worm classification. *Int. J. Comput. Sci. Inf. Secur.* **10**(4), 21–24 (2012)
22. C. Obimbo, A. Speller, K. Myers, A. Burke, M. Blatz, Internet worms and the weakest link: Human error, in *International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018
23. C. Zhang, S. Zhou, B. Chain, Hybrid epidemics—a case study on computer worm Conficker. *PLoS One* **10**(5), e0127478 (2015)
24. T. Mokoena, T. Zuva, Malware analysis and detection in enterprise systems, in *IEEE International Symposium on Parallel and Distributed Processing with Applications and IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, 2017
25. S.M.A. Sulieman, Y.A. Fadlalla, Detecting zero-day polymorphic worm: A review, in *21st Saudi Computer Society National Computer Conference (NCC)*, 2018
26. B. Wanswett, H.K. Kalita, The threat of obfuscated zero day polymorphic malwares: An analysis, in *International Conference on Computational Intelligence and Communication Networks (CICN)*, 2015
27. P. Vinod, Unknown metamorphic malware detection: Modelling with fewer relevant features and robust feature selection techniques. *IAENG Int. J. Comput. Sci.* **42**(2), 1–13 (2015)
28. D. Kiwia, A. Dehghantanha, K.-K.R. Choo, J. Slaughter, A cyber kill chain based taxonomy of banking Trojans for evolutionary computational intelligence. *J. Comput. Sci.* **27**, 394–409 (2018)
29. S.W. Soliman, M.A. Sobh, A.M. Bahaa-Eldin, Taxonomy of malware analysis in the IoT, in *12th International Conference on Computer Engineering and Systems (ICCES)*, 2017

Conceptual Model of Security Variables in Wi-Fi Wireless Networks: Review

Lorena Galeazzi, Cristian Barría, and Julio Hurtado

Abstract

Systems, data, users, and networks are essential in terms of information security. Systems, data, users and networks are essential in terms of information security. Wi-Fi wireless networks play a crucial role in increasing connectivity, as well as preventing and monitoring unauthorized access. Nonetheless, Wi-Fi wireless networks' security is conditioned by different variables incorporated in standards, norms, good practices, and various investigations concerning this topic.

In this way, the present research exposes an information survey of certain variables based on a narrative review, which allows their identification and possibly the incorporation of others. The results obtained from the survey will be shown through a conceptual model, which allows visualizing the different aspects required in the security that is applied to this technology.

Keywords

Wireless · Variables · Security · Wi-Fi · Standard · Controls · Best practices · Vulnerability · Multivariables · Wireless checklist

8.1 Introduction

New technologies associated with the usage of the internet for different applications have generated the need for improvements and optimization in the wireless networks, together with new features in speed, bandwidth, and communications [1, 2]. Security aspects have also been incorporated at the connection level [3]. They must counteract the vulnerabilities of Wi-Fi wireless networks and require direct intervention or interaction with the signal on the electromagnetic spectrum, with the particularity of being an unguided medium that makes protection even more difficult [4].

The possibilities of successfully accessing wireless networks depend on several factors that have been evolved over time, such as the level of the signal, passwords, their administration, and the recognition phase of the signals. Therefore, a series of specific tests related to security have been carried out that allowed establishing metrics in order to evaluate the behavior of both users and IT administrators, identifying parameters such as the percentages of users who update their devices and the type of encryption protocol used the most [5].

Currently, we rely on standards and the best practices, which provide us with a guide to audit, control, and strengthen Wi-Fi wireless networks. Each of the alternatives is possible to adjust to our network depending on the size of the organization, the level of criticality of the network, and the data that must be protected [6].

L. Galeazzi (✉)

Universidad Mayor, Santiago, Chile

Universidad del Cauca, Popayán, Colombia

e-mail: lorena.galeazzi@mayor.cl; lorenagaleazzi@unicauca.edu.co

C. Barría

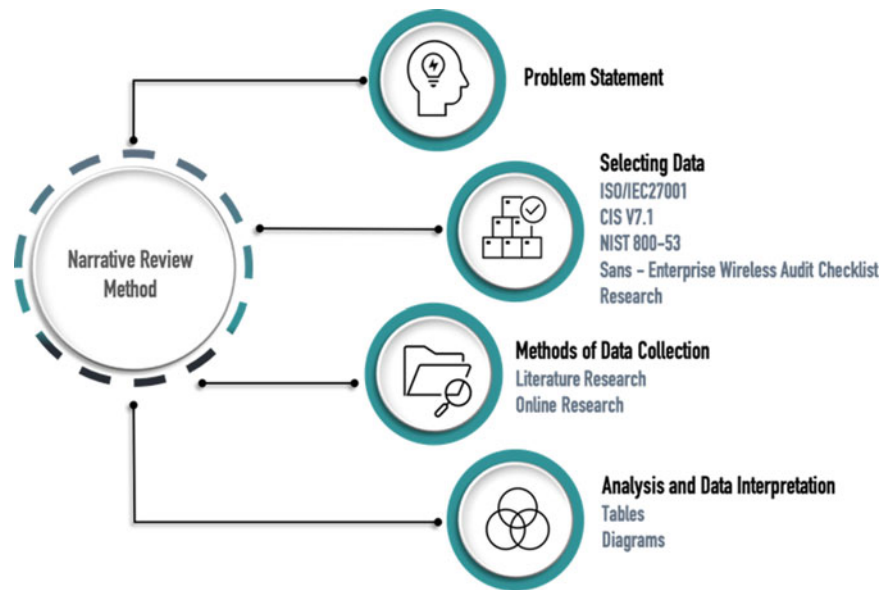
Universidad Mayor, Santiago, Chile

e-mail: Cristian.barría@mayor.cl

J. Hurtado

Universidad del Cauca, Popayán, Colombia

e-mail: ahurtado@unicauca.edu.co

Fig. 8.1 Phases of research

8.2 Problem Statement

Security in wireless networks is conditioned to multivariable [1]. All of them are results of studies, standards, the best practices, and information provided by literature on wireless pentesting. The number of aspects is to be added, depending on the scenario or reality in which there is a wireless network implementation [3].

The purpose of a conceptual model is to represent the important data of this research, allowing the identification of common and cross-sectional variables found in the reviewed literature.

It is necessary to have a conceptual model that clarifies the most important multivariables that may become the basis of security in a wireless network.

8.3 Scope

The scope of this investigation includes a review of the documentation, considered as of security in Wi-Fi wireless networks, the standards, ISO/EIC 27001, NIST 800-53, CIS controls, SANS-Enterprise Wireless Checklist, investigations and literature associated with security in Wireless Networks published between 2015 and 2020.

8.4 Study Methods

A narrative review methodology will be used to perform the analysis and interpretation of scientific literature. The main purpose is to use a process of search, data extraction, and

analysis based on an established method [7], as shown in Fig. 8.1.

8.5 Selecting Data

According to the magnitude of information related to security, it is important to generate a review of different standards, the best practices, controls, and investigations to define which variables are essential within security in wireless networks, as detailed below [3].

Regarding the selection of variables used in this research, a previous work published at the 9th International Congress, Witcom 2020, Mexico [3] was considered as the basis to continue the current study.

8.5.1 Standards

The standards are defined as a model, norm, pattern, or reference [8]. This type of information is essential when looking for a security reference both for auditing and optimizing the operation of an organization.

8.5.1.1 ISO/IEC 27001 Information Security Management

This standard specifies how to establish or implement an information security management system in an organization, in addition to contemplating the continuous improvement of processes within its purposes [9]. For this reason, the variables that can be associated with security in Wi-Fi wireless networks will be taken into consideration, as shown in the following Table 8.1.

Table 8.1 Standard security variables

Standard	Variables
ISO 27001	Segregation Cryptographic Network control Inventory Acceptable use of assets users
NIST 800-53	Security access Restrict ability to attack Protect external and internal electromagnetic interference cryptographic Imitative or manipulative communications deception signal parameter identification Firewall IDS Wireless monitoring Timing interactions Identify and explicitly authorize users Antenna Power levels Authorize Monitor Control the use
CIS controls	Firewall IDS Wireless access points Cryptography Multi-factor authentication Segregation Deploy port level access control Inventory Detect Disable wireless access Limit wireless access Disable capabilities Penetration testing

8.5.1.2 National Institute of Standards and Technology (NIST) 800-53

This standard provides security controls and evaluation procedures defined as recommendations for government organizations and information systems in the United States [10]. The associated variables are listed in the following Table 8.1.

8.5.1.3 Center for Internet Security (CIS) Controls

These controls are carried out by a non-profit global community of IT professionals, which continuously improves standards by providing the necessary tools for the prevention, protection, and incidents response to cyber threats for government entities in the United States [11]. The collected variables are declared in the following Table 8.1.

8.5.2 The Best Practices

The best practices are guidelines created from experience or intervention, which have been implemented positively [12].

Table 8.2 Security variables Sans – Enterprise wireless audit checklist

Controls	Variables
Sans Enterprise wireless audit checklist	Policy Firewall IDS/IPS Configuration management Password Cryptographic Configuration settings End user training SSID construction Session timeout Client isolation Radius server security Logging & monitoring

Sans – Enterprise Wireless Audit Checklist

As a definition of the best practices, they are controls tested and approved by many organizations worldwide, which are applied within a framework of reference. The SANS Institute is a cooperative research and education organization that works with other organizations to help the information security community. SANS makes 20 the best practices controls available to raise the level of maturity and control available for organizations, evaluating issues from management, architecture, configuration, awareness, encryption [12], which are listed in the following Table 8.2.

8.5.3 Research

Research related to security in wireless networks also describes the variables under study; therefore, a collection of information was done to carry out a survey from different studies.

The sources used to develop the variables study survey are listed in the following Table 8.3.

8.5.4 Books that Make References to Wireless Network Security

The books are also guides to the knowledge of wireless networks, as well as to security, management, and the best practices; therefore, variables have also been collected to complement the information for later analysis in this research, according to the following Table 8.4.

8.6 Analysis and Data Interpretation

The analysis of the data interpretation obtained confirms the presence of multivariables that obey different characteristics and functionalities; consequently, a search and identification

Table 8.3 Research security variables

Year	Author	Title	Variables
2015	Monsalve [13]	Security analysis of a WLAN network sample in Tunja, Boyacá, Colombia.	Cryptography Mac address Static IP address
2015	Mendez [14]	WEP, WPA and WPA2 encryption protocols vulnerability on wireless networks with Linux platform	Cryptography Attacks tools
2016	Prastavana [15]	Wireless security using Wi-Fi protected access 2 (WPA2)	Cryptography Eavesdropping Attacks types
2017	Gupta [16]	Ethical hacking and hacking attacks.	Methodology attack Attacks types
2017	Kalnins [17]	Security evaluation of wireless network access points	Cryptography WPS vulnerability SSID MAC address Social engineering attack
2018	Vanhoef [18]	Operating channel validation: preventing multi-channel man-in-the-middle attacks against protected Wi-Fi networks	Type attacks Cryptography 4-way handshake Monitoring (CSAs) Influence timing measurements (FTM)
2018	Vanhoef [19]	Release the Kraken: New KRACKs in the 802.11 standard	Security protocols 4-way handshake Type attacks MAC address Android, macOS, and OpenBSD
2018	Zhou [6]	IEEE 802.11ay based mmWave WLANs: Design challenges and solutions	SU – MIMO MU – MIMO MAC address Long distance Signal attenuation mmWave
2018	János [20]	Effects of the WPA2 KRACK attack in real environment	Wireless protocol Cryptography 4-way handshake Password Identify users Firmware update Manufacturers (publish updates) Operating system Antenna Wardriving
2018	Tchakounte [21]	Recognizing illegitimate access points based on static features: a case study in a campus Wi-Fi network	Wardriving Mac address WPS attacks Attacks types Cryptography Channels Security type SSID BSSID Architecture Antenna Software attacks Signal information
2019	Mahabub [22]	A voting approach of modulation classification for wireless network	Radio signal Modulation Signal Recognition spectrum

(continued)

Table 8.3 (continued)

Year	Author	Title	Variables
2019	Valchanov [23]	An empirical study of wireless security in city environment	Cryptography WPS attack Wardriving
2019	Sombatruang [24]	Factors influencing users to use unsecured Wi-Fi networks: Evidence in the wild	Unsecured Wi-Fi User age User education User income Battery power preservation heuristic.
2020	Kissi [25]	Penetration testing of IEEE 802.11 encryption protocols using Kali Linux hacking tools	Wardriving Penetration testing Cryptography Attacks types Radio frequency Passive and active attack 4-way handshake

Table 8.4 Security variables reference books

Year	Author	Title	Variables
2015	Ramachandran [26]	Kali Linux wireless penetration testing Beginner's guide	Cryptography SSID WPS Wireless attacks tools 4-way handshake HASH Radius server Client Active scanning Frequency analysis Wireless adapters Chipset Antennas Operative system Monitor mode
2015	Wright [27]	Hacking exposed wireless	SSID MIMO Radio frequency Physical layer Mac address Architecture
2016	Beard [28]	Wireless communication networks and system	Cryptography Mac address Static IP address Pentesting Attacks types
2016	Norman [29]	Computer hacking beginners guide	Cryptography Software Mode monitor Wireless adapters Chipset Password Mac address Static IP address Pentesting Attacks types

(continued)

Table 8.4 (continued)

Year	Author	Title	Variables
2018	Osterhage [30]	Wireless network security	Wardriving Hardware Software Data Eavesdropping Attacks types. Radio frequency Connection control Mac address Antenna WPS Access point Bandwidth Range Channel Wireless routers Architecture SSID Wireless adapters MAC address
2020	Yang [31]	Advanced wireless transmission technologies	MIMO Channel Energy harvesting
2020	Shen [32]	Encyclopedia of wireless networks	Wireless authentication MIMO Architecture
2020	Hoffman [33]	Wireless hacking with Kali Linux	Pentesting Pentesting tools Wireless adapter Passwords Cryptography Reconnaissance Attack types Physical security 4-way handshake Monitor mode Dictionary attacks Radius
2020	Haupt [34]	Wireless communications systems	Bandwidth Signal level Noise and interference Antenna Channel NIC Mac address Wireless router SSID Attacks types Cryptography Firewall IDS Radio frequency Software



Fig. 8.2 Related variables

were carried out, obtaining a more specific vision of each one of them.

To make a relationship conceptually, it was necessary to use a Venn diagram, so it was possible to determine the relationship of the variables obtained through the review performed (Fig. 8.2).

In the beginning, selecting variables seeks to find which variable stands out or is mostly considered in the literature. The following image shows the variables identified by each group concerning their source of origin.

It is possible to identify that three variables stand out as the most mentioned in the three groups of variables information sources.

Concerning the aspects raised in the previous points, it was possible to identify three variables, which have been mentioned and associated with the security of Wi-Fi wireless networks.

At first glance, there is a variable that has not been considered as such, but it is mentioned in studies, standards, and the best practices. This variable is associated with a series of vulnerabilities and threats, called risk, but has not been contemplated.

Therefore, as a result of this study and to be considered in the conceptual model, the variable “USER” will be established as the main variable, such as the three determined in the result of the analysis accomplished in this review.

In this regard, the literature speaks to us from different perspectives, as an attacker, as an information security officer, or as an incident response analyst. All this information is related to the vulnerability and the users in a complementary way.

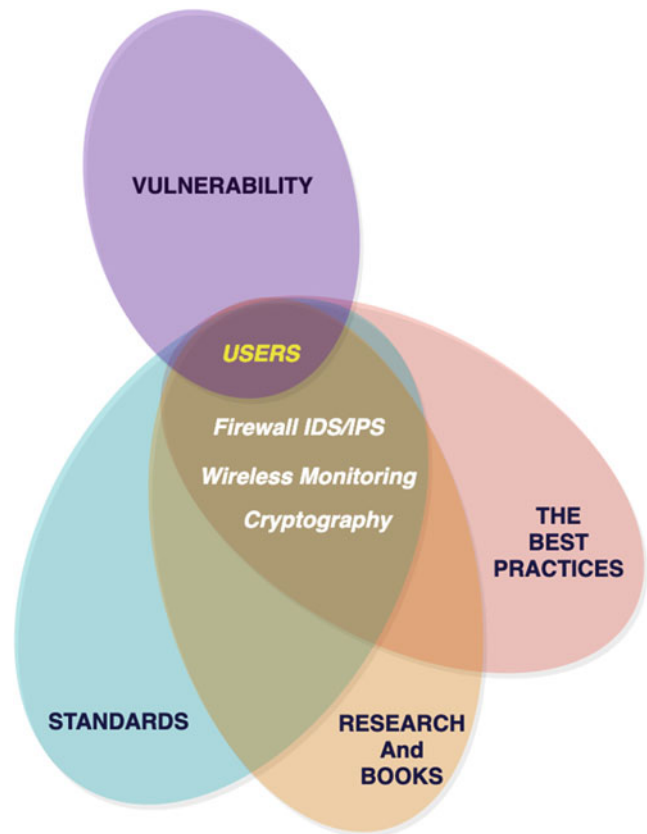


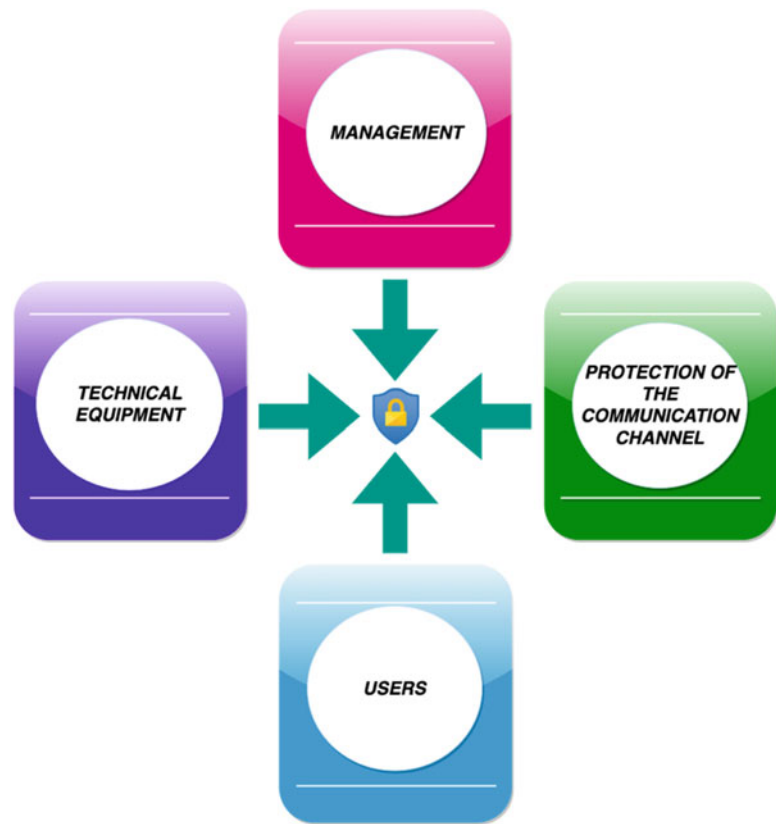
Fig. 8.3 Variables result

Besides, as part of this review, it shows us that there is no direct focus on users; therefore, it is vital to incorporate it as a variable to the conceptual model since users are classified as the weakest link in the chain [35] since the most significant part of the variables associated with vulnerabilities or threats are based on them, considering that technology, standards, and the best practices remain in continuous improvement and updating. However, users lack the same capabilities or continuous improvement.

Users directly relate to the storage, processing, and transfer of the information they handle [35].

On the other hand, the effectiveness of tools, platforms, applications, and the software will depend on a human; therefore, there are errors, negligence, bad practices, which increase the risk and vulnerability of organizations [36].

Likewise, there is literature available to predict users’ susceptibility to threats [37], in the same way at the level of passwords, where it is recommended that the complex one will help us increase security. Studies indicate that more than 80% of users create them using personal information [38]. Finally, everything indicates that this variable should be considered in all aspects, as shown in Fig. 8.3.

Fig. 8.4 Conceptual model

The conceptual model of security variables in Wi-Fi wireless networks is established by four concepts, which have been determined by the variables that compose it: Management, Technical Equipment, Protection of the Communication Channel, and End Users, the way it is shown in Fig. 8.4.

8.7 Conclusion

The application of a Narrative Review, being a type of review that allows the extraction of data from the literature, facilitates the identification regarding to what has been published on this subject. This review has a more selective way by applying data interpretation techniques, searching for patterns and trends among other aspects.

Regarding the determination of variables, it was carried out that the standards respond to the concept of information security and procedures for attacks on Wi-Fi wireless networks, based on the previous research.

Finally, the development of a conceptual model, when applying a Venn Diagram, allowed to group the identified variables and at the same time to be able to interpret graphically how the variables and the common points (intersections) between those that consider Security in wireless networks Wi-Fi are related.

8.8 Future Work

Further investigation is needed to perform an analysis, integration, and validation of new variables to be added to the conceptual security model in Wi-Fi wireless technology, that allows to have new tools at the security level in support of existing standards and good practices.

References

1. C. Barria, D. Cordero, L. Galeazzi, A. Acuña, Proposal of a multi-standard model for measuring maturity business levels with reference to information security standards and controls, in *Intelligent Methods in Computing, Communications and Control*, (Springer, Cham, 2020)
2. S. De Haes, W. Van Grembergen, *Enterprise Governance of Information Technology*, 2nd edn. (Springer, Belgium, 2015)
3. L. Galeazzi, C. Barria, J. Hurtado, A review of the security information controls in wireless networks Wi-Fi, in *9th International Congress, Witcom*, (Springer, Cham, 2020)
4. K. Erickson, *Networking Hacking* (Independently Published, United States, 2019)
5. H. Valchanov, J. Edikyan, V. Aleksieva, An empirical study of wireless security in city environment, in *Proceedings of the 9th Balkan Conference on Informatics*, 2019
6. P. Zhou et al., IEEE 802.11ay-Based mmWave WLANs: Design challenges and solutions. *IEEE Commun. Surv. Tutor.*

- 20(3), 1654–1681, third quarter 2018. <https://doi.org/10.1109/COMST.2018.2816920>
7. G. Paré, M. Trudel, M. Jaana, S. Kitsiou, Synthesizing information systems knowledge: A typology of literature reviews. *Inf. Manag.* **52**(2), 183–199 (2015)
 8. Rae, <https://dle.rae.es/estandar>. Last accessed 2020/10/02
 9. ISO, <https://www.iso.org/about-us.html>. Last accessed 2020/10/02
 10. NIST, <https://www.nist.gov/800-53>. Last accessed 2020/10/02
 11. CIS, <https://www.cisecurity.org/about-us/>. Last accessed 2020/10/02
 12. SANS, <https://www.sans.org/media/score/checklists/EnterpriseWirelessNetworkAudit.pdf>. Last accessed 2020/10/02
 13. J.A. Monsalve-Pulido, F.A. Aponte-Novoa, F. Chaparro-Becerra, Análisis de seguridad de una muestra de redes WLAN en la ciudad de Tunja, Boyacá, Colombia. *DYNA* **82**(189), 226–232 (2015)
 14. W.A. Méndez Moreno, D.J. Mosquera Palacios, E. Rivas Trujillo, WEP, WPA and WPA2 encryption protocols vulnerability on wireless networks with Linux platform. *Rev. Tecnura* **19**, 79–87 (2015). <https://doi.org/10.14483/udistrital.jour.tecnura.2015.SE1.a06>
 15. Prastavana et al., Wireless security using Wi-Fi protected access 2 (WPA2). *Int. J. Sci. Eng. Appl. Sci.* **2**(1), 374–382 (2016)
 16. A. Gupta, A.A. Student, Ethical hacking and hacking attacks. *Int. J. Eng. Comput. Sci.* **6**(6), 2319–2324 (2017). <https://doi.org/10.18535/ijecs/v6i4.42>
 17. R. Kalniņš, J. Purins, G. Alksnis, Security evaluation of wireless network access points. *Appl. Comput. Syst.* **21**(1), 38–45 (2017)
 18. M. Vanhoef, N. Bhandaru, T. Derham, I. Ouzieli, F. Piessens, Operating channel validation: Preventing multi-channel man-in-the-middle attacks against protected Wi-Fi networks, in *WiSec*, 2018
 19. M. Vanhoef, F. Piessens, Release the Kraken: New KRACKs in the 802.11 standard, in *CCS*, 2018
 20. D.J. Fehér, B. Sandor, Effects of the WPA2 KRACK attack in real environment, in *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, Subotica, 2018
 21. F. Tchakounte et al., Recognizing illegitimate access points based on static features: A case study in a campus Wifi network. *Int. J. Cyber Secur. Digit. Forensic.* **8**(4), 279–291 (2018)
 22. A. Mahabub, A. Sultan, A voting approach of modulation classification for wireless network, in *Proceedings of the 6th International Conference on Networking, Systems and Security (NSysS '19)*, (Association for Computing Machinery, New York, 2019), pp. 133–138
 23. H. Valchanov, J. Edikyan, V. Aleksieva, An empirical study of wireless security in city environment, in *Proceedings of the 9th Balkan Conference on Informatics (BCI'19)*, (Association for Computing Machinery, New York, 2019), pp. 1–4, Article 11
 24. N. Sombatrung, L. Onwuzurike, M. Sasse, M. Baddeley, Factors influencing users to use unsecured Wi-Fi networks: Evidence in the wild, in *Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '19)*, (Association for Computing Machinery, New York, 2019), pp. 203–213
 25. M. Kissi, M. Asante, Penetration testing of IEEE 802.11 encryption protocols using Kali Linux hacking tools. *Int. J. Comput. Appl.* **176**(32), 26–33 (2020)
 26. V. Ramachandran, C. Buchanan, *Kali Linux Wireless Penetration Testing Beginner's Guide* (Packt Publishing Ltd., Birmingham, 2015)
 27. J. Wright, J. Cache, *Hacking Exposed Wireless* (McGraw-Hill Education, New York, 2015)
 28. C. Beard, W. Stalling, *Wireless Communication Networks and System* (Pearson Higher Education, Inc, Hoboken, 2016)
 29. A. Norman, *Computer Hacking Beginners Guide* (CreateSpace Independent Publishing, North Charleston, 2016)
 30. W. Osterhage, *Wireless Network Security* (Taylor & Francis Group, Boca Raton, 2018)
 31. H. Yang, M. Alouini, *Advanced Wireless Transmission Technologies* (Cambridge University Press, New York, 2019)
 32. X. Shen, X. Zhang, *Encyclopedia of Wireless Networks* (Springer Nature Switzerland AG, Cham, 2020)
 33. H. Hoffman, *Wireless Hacking with Kali Linux: Learn Fast How to Hack Any Wireless Networks Penetration Testing Implementation Guide* (Independently Published, 2020)
 34. R. Haupt, *Wireless Communications Systems* (John Wiley & Sons, Inc, Hoboken, 2020)
 35. SANS, <https://www.sans.org/security-awareness-training/blog/why-human-weakest-link>. Last accessed 2020/10/02
 36. G.L. Orgill, G.W. Romney, M.G. Bailey, P.M. Orgill, The urgency for effective user privacy-education to counter social engineering attacks on secure computer systems, in *Proceedings of the 5th Conference on Information Technology Education (CITC5 '04)*, (Association for Computing Machinery, New York, 2004), pp. 177–181
 37. R. Heartfield, G. Loukas, D. Gan, You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks. *IEEE Access* **4**, 6910–6928 (2016)
 38. J. Still, A. Cain, Over-the-shoulder attack resistant graphical authentication schemes impact on working memory, in *AHFE 2019, AISC 960*, ed. by T. Ahram, W. Karwowski, (Springer Nature Switzerland AG, Cham, 2020), pp. 79–86

Cybersecurity Analysis in Nodes that Work on the DICOM Protocol, a Case Study

David Cordero and Cristian Barría

Abstract

Currently the Internet is the main tool for interconnection of systems and data processing, SCADA systems, marketing, government; Even medical systems, among others, need to be connected to the internet to facilitate the development and processing of their data, the services that operate with the DICOM protocol (Digital Imaging and Communications in Medicine) work with medical equipment. This protocol is the universal format for the exchange of medical images, due to which it is used worldwide for communication between devices, the so-called PACS servers (Image Archiving and Communication System). These are information receptacles where medical centers store X-rays, files, personal information of patients, information of the treating physician, among others. The purpose of this research is to carry out a cybersecurity analysis in the operational and connected nodes in Chile that operate with the DICOM protocol among their services, with the execution of a modified experimental design that will allow the discovery of active nodes, discovery of exposed services and vulnerabilities, the analysis of said services as well as their vulnerabilities, their categorization and finally the validation of the vulnerabilities found. It seeks to know the current situation in cybersecurity issues of the nodes that use the DICOM protocol for communication, identifying the possible attack vectors that third parties may use in order to compromise the integrity, confidentiality, availability and authenticity of said systems.

Keywords

Protocol · DICOM · Servers · PACS · Threats · Medical · Critical · Infrastructure · Nodes · Vulnerability

9.1 Introduction

In an interconnected world, the interaction between services is essential, which facilitate access to repositories, operation between specialized machinery, information systems, government systems, marketing systems, even SCADA systems (Supervisory Control and Data Acquisition) [1]. These systems are used in critical infrastructure, all connected to each other through the Internet, Medical services are not alien to this need for interconnection, since medical equipment communicates with each other to send data and information, radiographic systems or repositories of exams, all of which use different communication protocols to carry out the transfer of information between them. One of the most used is the DICOM (Digital Imaging and Communications in Medicine) protocol [2], this protocol is the universal format for the exchange of medical images and allows the transfer of information, the reading and saving of examinations, files, images among others. The so-called PACS servers (Picture Archiving and Communication System) [3] use this protocol for communication between services, to establish a client-server mode for nodes that connect and require information from said repositories, however this interaction leaves the services exposed to different threats since they are connected through the internet.

The objective of this research is to carry out an analysis focused on the cybersecurity of the nodes available on the internet that use the DICOM protocol and that these operate in the context of Chile.

D. Cordero (✉) · C. Barría
Centro de Investigación de la Universidad Mayor CICS, Universidad Mayor, Santiago, Chile
e-mail: david.cordero@mayor.cl; cristian.barría@umayor.cl

In relation to the above, there are publications that account for the threats in said protocol, such as the one published in early 2020 by Wang [4] that exposes flaws found in the PACS servers and that allow the denial of services (availability of the compromised system) From automated software tests, providing invalid, unexpected or random data (fuzzing). Private companies also make reports on the current situation of nodes that work on this protocol, delivering comprehensive results of the number of examinations exposed in said nodes. The private organization Greenbone [5] conducted a study in 2019 publishing a list of vulnerable nodes that may be compromised in the confidentiality of their resources, including study publications that present the level of security in PACS servers implemented in hospitals, demonstrating the data that can be obtained through these services and the lack of digital signatures in modifying said information [6].

Due to the situation in which countries find themselves due to the current COVID-19 pandemic, which puts all the world's medical systems to the test, an attack on these could cause a direct impact on the population, since both the denial of services (availability), such as the integrity and authenticity of the stored data, and the confidentiality of these are critical in the area, and although the present study will only cover Chile as a country, however this experiment can be applied in any region of the world. This work is based on the research on the cybersecurity analysis of PACS-DICOM servers in Chile [7] that exposes the development of a cybersecurity analysis on an experimental design, which includes the recognition, analysis, diagnosis and evidence of the operating nodes, covering only the confidentiality of data to which said nodes are exposed from queries through the DICOM protocol, using methods for obtaining information such as WADO (Web Access to DICOM Object) which allows the interaction of the PACS server is carried out through the web and in a standard way [8], this method is typical of the DICOM protocol and from default or deficient configurations, it was possible to obtain medical images, pretending to be a PACS-viewers client, of such a way of showing data exposure (data response without access protection) covering only the queries from the default ports in said protocols (104 and 11,112).

For the development of this research, the exposed nodes are identified from a range of dates that ranges from June, July, August and September, in which the information of active nodes is collected, which are subjected to a cybersecurity analysis In this way, considering the services and open ports of each identified node, as well as the vulnerabilities detected, contrasting said information from a database of vulnerabilities published on cve.mitre.org. CVE (Common Vulnerabilities and Exposures) [9] allows to relate a known vulnerability with a unique identification number (CVE-ID) that describes said vulnerability, identifies the software versions that are affected and their possible mitigation so-

lutions, among others, determining in this way the active vulnerabilities of PACS-DICOM servers exposed to internet threats.

The reformed experimental design proposed for the execution of this research is based on work [7] with a new scope in the discovery stage, since it covers a broader spectrum in possible attack vectors, registering the services and exposed versions of each node in isolation.

In the first instance, each version of some software, service or protocol discovered will be checked against the CVE database. It should be noted that this process is automated from a script that consults this information, which can be an easy process for third parties to execute.

The present investigation is developed in the following way: (1) presentation of previous works and their design changes in the experiment process; (2) description of each process in the reformulated and new scope design; (3) description of the data that will be presented later; (4) evaluation of the data found in this research to finally draw conclusions.

9.2 Related Work and Reformulation of Experimental Design

The experimental design described in the research previously carried out and in which said procedure included four processes for its development (data discovery, categorization, analysis and validation) as shown in Fig. 9.1, was used as the basis for the execution of an analysis with a greater research scope, based on an experimental process redesigned to cover the vulnerabilities present in each identified node. This makes it possible to cover a broader spectrum and a more effective analysis of the cybersecurity of the identified systems, thus covering the information security quartet [10], recognizing the vulnerabilities that may compromise, (1) availability: ability to guarantee that both the system and the data are available at all times, (2) confidentiality: ownership of the information that guarantees that only authorized individuals or organizations can access it, (3) integrity: correctness and completeness of the data in a database or repository of the information residing in each node, and (4) authenticity: which is related to the veracity of the data, said quartet is displayed in the following Fig. 9.2.

To meet this new scope, new processes are added that will allow the collection of data samples from the discovery of services and potential vulnerabilities that can compromise a node that is working under the DICOM protocol (PACS servers).

Said reformulation includes the development of node discovery, discovery of active services and versions of func-

Fig. 9.1 Experimental design [7]

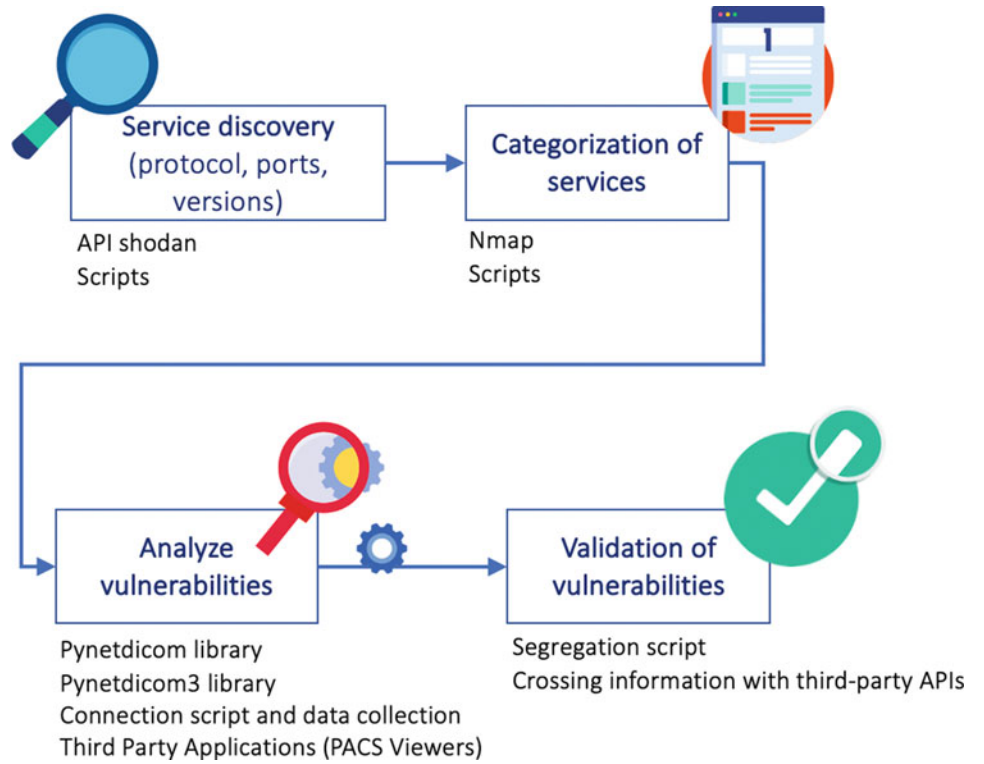
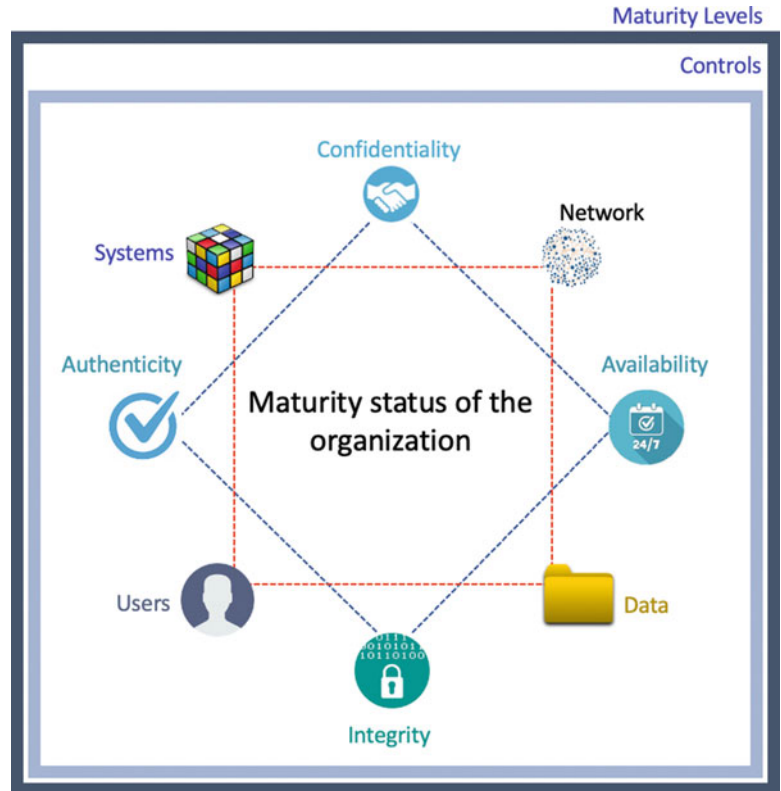


Fig. 9.2 Information security quartet on the state of maturity of an organization [10]



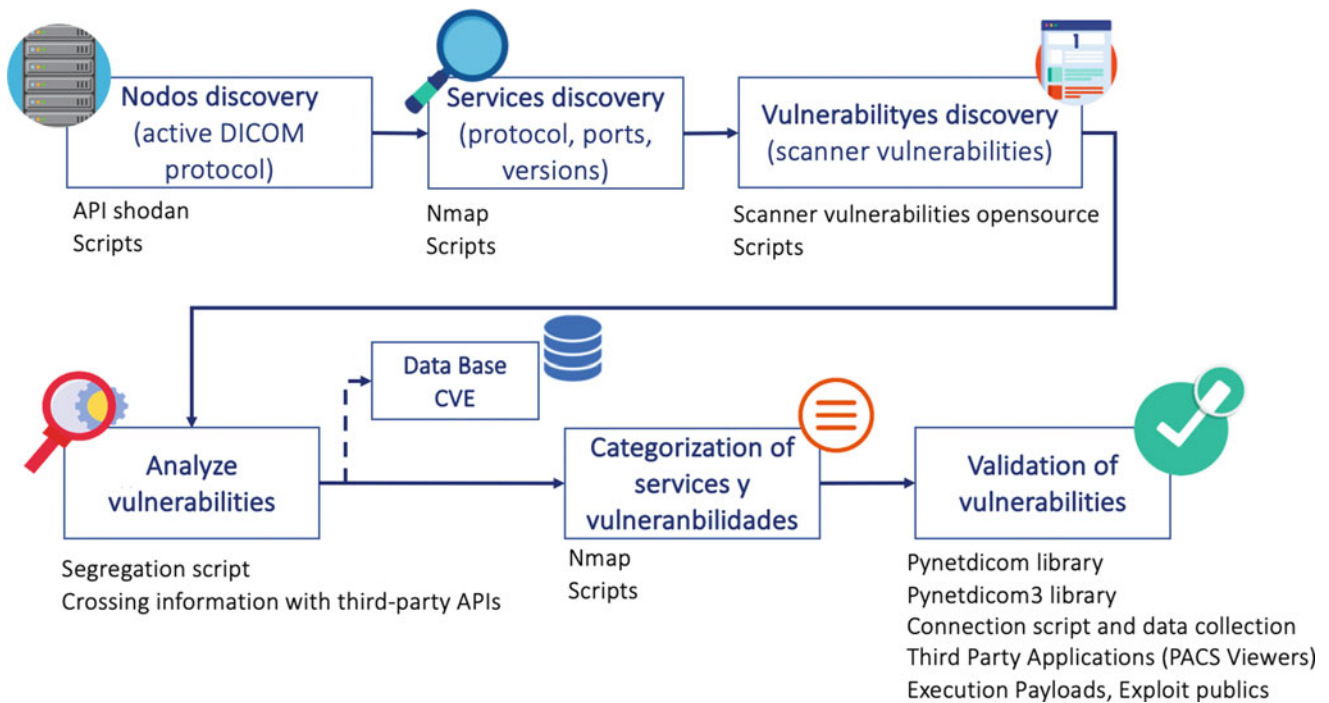


Fig. 9.3 Reformulated experimental design

tional systems, the discovery of vulnerabilities linked to said services, analysis of services and vulnerabilities referenced from a database for public use, which is in charge of to register each one of them with a unique code, thus identifying each vulnerability individually, then the categorization of services and vulnerabilities and finally the validation of vulnerabilities found, which will allow the measurement of success rates of an attack on said platforms and systems. This reformulated experimental design is seen in Fig. 9.3.

9.3 Experimental Design for Classification of Nodes

After the reformulation of the experimental design, a chain of processes is carried out that will allow the classification and current situation of each node, allowing to visualize the specific region studied.

- Discovery of nodes:** The first process of experimental design is the discovery of nodes. Here the first stage of the previously exposed design is redefined with the support of the Shodan API [11], the nodes that use the DICOM protocol for their communications are identified, thus obtaining a list of active servers, this discovery of nodes, it is executed and monitored between the months of June, July, August and September of the year 2020, since there are many nodes that are only operational for a few

hours a day and at certain times of the week. From this data collection it is possible to identify and work in isolation each node that uses the DICOM protocol (PACS servers) and perform a subsequent analysis on them.

- Discovery of services:** During this experimental design process, each node will be subject to an analysis of active services, thus allowing to recognize the different software and protocols that are operating in each node, as well as the versions of software used. For the port scanning, the nmap tool will be used, an open source program that is used to scan and identify ports and software services (the complete scan configuration will be used considering the 65,535 tcp and udp ports), said use of nmap will be supported through plugins to support vulnerability scans.
- Discovery of vulnerabilities:** The vulnerability discovery process will be released after each node is subjected to a vulnerability scan. This procedure will be carried out from a free code vulnerability scanner available in any version of Kali Linux [12] (distribution based on GNU/Linux designed for auditing and computer security).
- Vulnerability analysis:** During this vulnerability analysis process, each node is subjected to a vulnerability analysis based on its active services, said analysis is carried out from an open source scanner and each finding is compared with the cve.mitre.org database (CVE), an organization that collects, categorizes and makes known vulnerabili-

ties available to the world, thus identifying whether the versions of software that are implemented in the node under analysis can be used by third parties based on the technical detail of the scope of exploitation of vulnerability.

5. **Categorization of services and vulnerabilities:** From the information collected by the processes of discovery of services and vulnerabilities, and their analysis, the software and versions active in the nodes are identified and compared with the database of cve.mitre.org. This allows obtaining technical details of the scope of the exploitation of the vulnerability considering the severity, complexity and impact of these on the vulnerable system, as well as a classification of said services based on their functionality in the affected node's system.
6. **Validation of vulnerabilities:** The vulnerability validation process is fed by the information obtained in the previous processes of discovery, analysis and categorization of services and vulnerabilities, after which procedures are executed to obtain data in the operating nodes that have vulnerabilities and exploitable. Queries about the DICOM protocol are made to each node, using methods to establish connection and obtain information to the node, using libraries built in Python and for public use, such as Pynetdicom and Pynetdicom3, in order to build a custom PACS-viewers client for this case. It is also possible to access this data with third-party PACS-viewers – you only need the node address and query with a functional method of the DICOM protocol (WADO) of obtaining on the available port of the service. The nodes from which the extraction of information is possible will be considered as vulnerable, since the breach of confidentiality in these servers is evident, as well as the nodes that have serious vulnerabilities (high impact on the integrity, confidentiality and availability of the system) and have a low level of exploitation, either because there is an exploit or public code that allows the exploitation of said vulnerability. It should be noted that no organization is linked to the exposure of your data, this work is only carried out for research purposes in such a way as to make known the current situation in the region of Chile before the DICOM services and how it is presented before cyber threats on the internet.

allow taking a sample for the correct cybersecurity analysis in each node.

Servers or operational nodes: They correspond to the servers that have activity in the Chilean region and work with the DICOM protocol. This compilation was carried out in a period of 4 months (June, July, August and September).

Nodes Vulnerable to data extraction through the DICOM protocol: From these nodes it is possible to extract data using the dicom protocol (WADO) communication methods in such a way as to access information resident in the node, without access credentials or authentication methods in such a way as to violate the confidentiality of said node.

Servers with vulnerabilities of high severity and low complexity of execution: The nodes that have vulnerabilities registered in the cve.mitre.org CVE database and that are classified with a high severity since the exploitation can compromise the system in question both in its integrity, confidentiality and availability of its system and also have low complexity at the time of exploitation, either because there is public code or exploit that can be easily used by third parties.

From the obtaining of vulnerabilities listed in each node, they will be categorized according to their level of impact on the system and the number of vulnerabilities related to said level of impact:

1. **Severity:** This value is directly linked to the impact of the vulnerability, determining its level from the system compromise, and that in some way affects part of the information security triad (integrity, confidentiality, availability).
2. **Number of vulnerabilities:** Numerical quantity of vulnerabilities found in all the analyzed nodes and that cover the region of Chile.
3. **Number of different vulnerabilities in compromised nodes:** The vulnerabilities are counted centrally from the nodes, for example if two nodes have both two high vulnerabilities (same CVE) this value corresponds to four vulnerabilities since in practice they are the same vulnerabilities but on different nodes.

The categorization of vulnerabilities is understood from the active services that are operating in the nodes. The data collected shows which are the vulnerable services or software and their variables are described below.

1. **Associated service:** This variable is given by the name of the identified service, it does not include versions of these.

9.4 Description of the Data

The data obtained in each stage of the exposed experimental design make up variables from which the data will be obtained, these variables will be described below, and these will

2. **Number of vulnerabilities:** Numerical value given by the number of occurrences within the vulnerability analysis.
3. **Percentage of vulnerabilities:** Percentage value based on the number of occurrences of the service based on the number of vulnerabilities found in all nodes.

For the presentation of the current situation of the nodes that work on the DICOM protocol, only the vulnerabilities found classified as high severity and low-medium complexity were considered in this investigation, and that from their exploitation can currently compromise the PACS server systems DICOM in Chile, the variables used for this information are the following:

1. **CVE:** Common Vulnerabilities and Exposures corresponds to the unique identifier of the public database of cve.mitre.org, which lists known vulnerabilities and identifies them with a unique id.
2. **Severity:** As in the categorization of samples, this value is directly related to the impact of the vulnerability if it is exploited in the system. This severity is related to the ease of exploitation (complexity).
3. **Complexity:** Level necessary for the execution and exploitation of the vulnerability in question. It is linked to the existence of tools, algorithms or code available on the internet that facilitate the exploitation of the vulnerability.
4. **Impact:** Measurement of the level of compromise in the system, from the execution of a vulnerability, delimited by the triad of information security, availability, integrity and confidentiality.
5. **Number of vulnerable nodes:** Numerical value that relates the number of PACS-DICOM servers that are in Chile and that have vulnerabilities categorized in a high severity level and a low-medium complexity.

9.5 Evaluation of the Data

From the information collected, the current situation in cybersecurity issues of the nodes available in Chile that use the DICOM protocol for the exchange of medical images is evidenced, from the execution of the experimental design. The research allowed the identification of 45 active servers (nodes that connected at least once a week were considered active), from these nodes a direct query was carried out through the DICOM protocol using the WADO method on standard ports of said protocol (104 and 11,112) without use of credentials. Through the execution of this procedure, a total of 170,652 examinations linked to users (patients) of medical services were obtained, the nodes from which this information could be obtained added up to ten (10), as well as from the analysis Of the forty-five (45) servers, thirteen (13) of them were detected as having vulnerabilities

Table 9.1 General data of nodes working on pacs-dicom protocol

Condition	Servers
Servers or operational nodes	45
Nodes in which data can be extracted from the DICOM protocol without authentication through the WADO method (ports 104 and 11,112)	10
Servers with vulnerabilities of high severity and low-medium complexity of execution	13

Table 9.2 Classification by level of severity of vulnerabilities in operating nodes

Gravity	Number of vulnerabilities	Number of different vulnerabilities in vulnerable nodes
High	19	42
Medium	69	152
Low	8	14
Total	97	

classified as high severity and low-medium complexity of exploitation. Below in Table 9.1 the information described is displayed.

In the forty-five (45) active nodes, ninety-seven (97) vulnerabilities cataloged by cve.mitre.org CVE were identified, of which in a subsequent analysis they were classified as nineteen (19) of high severity, sixty-nine (69) of medium severity and eight (8) of low severity, all this classification directly related to the complexity of exploitation. The information is shown below in Table 9.2.

Within the analysis of the nineteen (19) high severity vulnerabilities, it is possible to determine their level of impact on the system that may be compromised, thus identifying that six (6) of these affect with a total compromise the integrity, confidentiality and system availability, twelve (12) with a partial compromise of the system and one (1) affecting the availability of the vulnerable node. Next, in Table 9.3, only the vulnerabilities identified by the unique id provided by cve.mitre.org, their level of severity, the complexity of exploitation, the impact of the vulnerability on the system and the number of occurrences within are detailed of the total list of nodes identified as operational.

From the identification of the ninety-seven (97) vulnerabilities, it is possible to break down the active services associated with each node, thus identifying the most affected of the total number of nodes, as shown in Table 9.4.

From the information obtained in Table 9.4 it is possible to perform a categorization depending on the functionality or provision to which each vulnerable service is associated. Figure 9.4 shows that sixty-six percent (66%) of the vulnerabilities are associated with web services technologies, twenty-seven percent (27%) with cryptographic communication services, and five percent (5%) with services. of remote connection to said vulnerable nodes, one percent (1%) to vul-

Table 9.3 CVE data identified as serious vulnerability in active nodes

CVE	Gravity	Complexity	Impact	Nodes
CVE-2019-0211	High	Low	Total commitment to the integrity, confidentiality and availability	7
CVE-2017-7679	High	Low	It partially affects the integrity, confidentiality and availability.	3
CVE-2017-7668	High	Low	It partially affects the integrity, confidentiality and availability	3
CVE-2017-3167	High	Low	It partially affects the integrity, confidentiality and availability	3
CVE-2017-3169	High	Low	It partially affects the integrity, confidentiality and availability	3
CVE-2010-2730	High	Medium	Total commitment to the integrity, confidentiality and availability	2
CVE-2010-3972	High	Low	Total commitment to the integrity, confidentiality and availability	2
CVE-2010-1256	High	Medium	Total commitment to the integrity, confidentiality and availability	2
CVE-2018-19518	High	Medium	Total commitment to the integrity, confidentiality and availability	2
CVE-2019-9020	High	Low	It partially affects the integrity, confidentiality and availability	2
CVE-2019-9021	High	Low	It partially affects the integrity, confidentiality and availability	2
CVE-2018-12882	High	Low	It partially affects the integrity, confidentiality and availability	2
CVE-2019-9023	High	Low	It partially affects the integrity, confidentiality and availability	2
CVE-2019-9641	High	Low	It partially affects the integrity, confidentiality and availability	2
CVE-2016-2177	High	Low	It partially affects the integrity, confidentiality and availability	1
CVE-2016-6303	High	Low	It partially affects the integrity, confidentiality and availability	1
CVE-2016-6304	High	Low	No impact on integrity and confidentiality, full compromise of system availability	1
CVE-2016-2182	High	Low	It partially affects the integrity, confidentiality and availability	1
CVE-2019-0708	High	Low	Total commitment to the integrity, confidentiality and availability	1

Table 9.4 Categorization of functionality related to serious vulnerabilities found in active nodes

Associated service	Number of vulnerabilities	Percentage of total vulnerabilities (%)
Apache HTTP server	35	36
PHP	28	29
OpenSSL	26	27
Open SSH	4	4
Remote desktop services	1	1
Procesadores SUSE	1	1
Microsoft FTP service	1	1
Lighttpd	1	1

nerabilities associated with the node's hardware and finally one percent (1%) to file transfer services.

% of the active nodes in Chile are exposed to serious vulnerabilities, while 78% of these nodes did not detect vulnerabilities that affect the integrity, availability, authenticity and confidentiality with reference to the published CVEs.

9.6 Conclusión

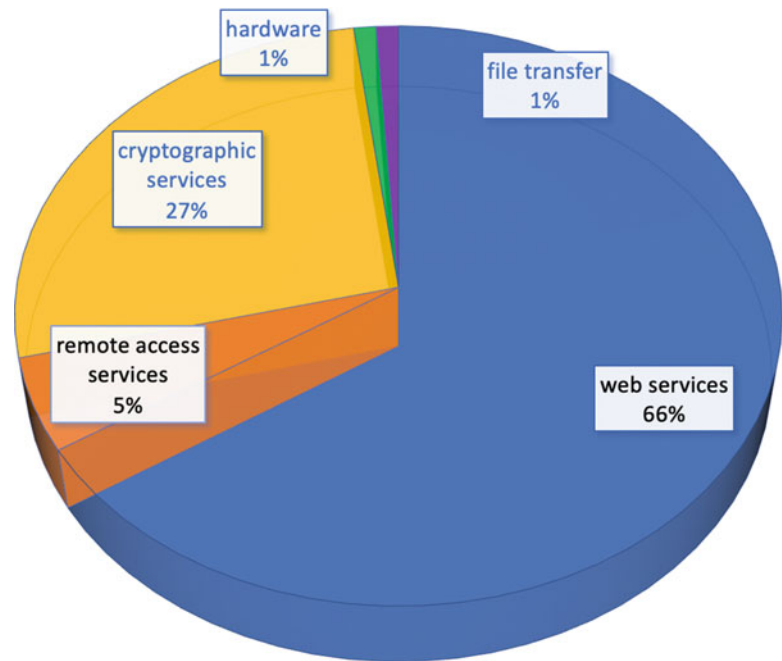
Due to the current situation in the world due to the pandemic caused by COVID-19 (SARS-CoV-2), medical services become even more fundamental and critical. The manipulation, denial or corruption in the integrity of these, directly affects

the population. The development of this research allowed to know the current situation that Chile is experiencing since the operation of these services, clearly identifying the percentage of nodes that can be compromised by third parties. Twenty-two percent (22%) of active servers in Chile present serious vulnerabilities that can be exploited and cause real problems for both the population and private organizations that have these services connected to the Internet.

Thus, it was also possible to identify exactly the services that were exposed to third-party attacks, such data is directly related to the initial configuration of each node and to the lack of update in the versions of the active services of each node. As evidenced in the research, sixty-six percent (66%) of the vulnerabilities found are associated with web services, being a server or technologies associated with its implementation. Such entry can allow a cyber attack on the organization, thus compromising the personal data of hundreds or thousands of patients.

The study highlights the constant need to keep systems up to date and protected against cyber threats, especially when today the use of ransomware by cybercriminals is very popular, which, based on these inputs, can cause serious problems that will eventually lead to be irreparable. This is aggravated by the current health crisis, which is why the correct operation of these medical services is essential, since the error in operations of these nodes can mean large monetary losses for organizations and affect the quality of life of part of the population.

Fig. 9.4 Categorization of vulnerable services in operational nodes in Chile



References

1. S. Samtani, S. Yu, H. Zhu, M. Patton, J. Matherly, Identifying supervisory control and data acquisition (SCADA) devices and their vulnerabilities on the Internet of Things (IoT): A text mining approach. *IEEE Intell. Syst.*, 63–73 (2018). <https://doi.org/10.1109/MIS.2018.111145022>
2. P. Leite, S. Carvalho, P. Teixeira, Á. Rocha, DICOM functionality assessment, in *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*, Lisbon, 2017, pp. 1–4, <https://doi.org/10.23919/CISTI.2017.7975888>
3. S. Cohen, F. Gilboa, U. Shani, *Proceedings Volume 4685, Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation* (2002). <https://doi.org/10.1117/12.467019>
4. Z. Wang, Q. Li, Q. Liu, B. Liu, J. Zhang, T. Yang, Q. Liu, DICOM-Fuzzer: Research on DICOM vulnerability mining based on Fuzzing technology, in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, (2020). https://doi.org/10.1007/978-3-030-41114-5_38
5. Greenbone Networks GmbH.: Information security report confidential patient data freely accessible on the internet. *Ciber Resilience Report* (2019)
6. A. Farhadi, M. Ahmadi, The information security needs in radiological information systems—an insight on state hospitals of Iran, 2012. *J. Digit. Imaging* **26**, 1040–1044 (2013). <https://doi.org/10.1007/s10278-013-9618-3>
7. D. Cordero, C. Barria, Cybersecurity analysis on PACS-DICOM servers in Chile, in *Witcom 2020 The 9th International Congress of Telematics & Computing*, 2020
8. G. Koutelakis, G. Triantafyllou, G. Mandellos, D. Karageorgopoulos, A web PACS architecture based on WADO service of DICOM standard (2005), pp. 284–288
9. [cve.mitre.org](https://www.cve.mitre.org/) [Online]. Available: <https://www.cve.mitre.org/>. Last access: October 12, 2020
10. C. Barria, D. Cordero, Proposal of a multi-standard model for measuring maturity business levels with reference to information security standards and controls, in *Intelligent Methods in Computing, Communications and Control. ICCCC 2020. Advances in Intelligent Systems and Computing*, ed. by I. Dzitac, S. Dzitac, F. Filip, J. Kacprzyk, M. J. Manolescu, H. Oros, vol. 1243, (Springer, Cham, 2020). https://doi.org/10.1007/978-3-030-53651-0_10
11. shodan [Online]. Available: <https://www.shodan.io/>. Last access: October 12, 2020
12. G. Najera-Gutierrez, J. Ansari, *Web Penetration Testing with Kali Linux: Explore the Methods and Tools of Ethical Hacking with Kali Linux*, 3rd edn. (Packt Publishing Limited, Birmingham, 2018), ISBN: 1788623800, 9781788623803

Robert Banks, Jim Jones, Noha Hazzazi, Pete Garcia,
and Russell Zimmermann

Abstract

Cybersecurity risk management often uses experience-based data to quantify the potential risks of new security technologies based on their exploitability and impact. However, use of such data may be limited and is rarely reusable because it often contains confidential information. This paper proposes a new approach using the Department of Homeland Security's public National Vulnerability Database (NVD) for information on known vulnerabilities, and MITRE's public Common Attack Pattern Enumeration and Classification (CAPECTM) tools as the basis of a risk scoring system.

Keywords

Bayesian belief network (BBN) · Common attack pattern enumeration and classification (CAPECTM) · Common vulnerability scoring system (CVSS) · Cyber

survivability endorsement (CSE) · Data framework · Generation of security · Linear regression · National Vulnerability Database (NVD) · Risk assessment · And sensitivity analysis

10.1 Introduction

Risk management is a useful tool for controlling risk, although it has limitations when trying to produce quantitatively accurate, reliable estimates of exploitability and impact.

Where experience-based data is required to quantify the exploitability and impact of potential security threats, such data is often limited and rarely reusable because it involves confidential data. Therefore, risk estimation would benefit from using publicly available data sources. We develop and apply a Bayesian Belief Network (BBN) to generate probabilities within this risk management system to assess the risk posed by new technologies.

This approach enables a more accurate and trustworthy way of quantitatively estimating the exploitability and impact of new technologies, based entirely on public data. We use MITRE's public Common Attack Pattern Enumeration and Classification (CAPEC) tools [2] and the Department of Homeland Security's public National Vulnerability Database (NVD) [10] to generate risk measurements of technology vulnerabilities using Bayesian Belief Networks (BBN). Our method produces accurate and trustworthy quantitative estimates of the exploitability and impact of new technologies based entirely on public data.

R. Banks (✉)

Volgenau School of Engineering, George Mason University, Fairfax, VA, USA
e-mail: rbanks3@gmu.edu

J. Jones

Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA
e-mail: jjonesu@gmu.edu

N. Hazzazi

Department of Electrical and Computer Science, Howard University, Washington, DC, USA
e-mail: Noha.Hazzazi@howard.edu

P. Garcia

PricewaterhouseCoopers, LLP, Miami, FL, USA
e-mail: Pete.Garcia@pwc.com

R. Zimmermann

SAIC, Chantilly, VA, USA
e-mail: Russell.F.Zimmermann@saic.com

The contributions of this methodology are three-fold:

1. A quantitative security risk level estimation for new security technologies using the CVSS for any computerized system's critical factor security.
2. Demonstration of the use of publicly available NVD data to estimate distributions used in the CVSS analysis. The CVSS uses the NVD data as input information to construct a model for analysis within a BBN structure.
3. A BBN model structure that provides flexibility for the model implementation. It allows other sources to be combined with estimates of the CVSS information and various levels of abstraction for input information.

10.2 Problem Context

Our national security operations need abilities beyond today's cybersecurity methods that focus on the goals of protect, detect, tolerate, and survive. Rather, these operations need industry partnerships and access to the newest technologies, but they also need objective assessment capabilities to make good risk-based decisions.

As businesses come to terms with the increasing threats in the cyber and cyber-physical spaces, instituting the right policies is critical to harness private sector capabilities. Due to ever changing risks and threats, the government needs to develop private sector support for establishing sound cybersecurity policy, while not creating regulations that hinder businesses more than help. For progress to be made, cybersecurity experts will need to view solutions in economic, policy, and risk terms, rather than just technology. This work provides a public data-based objective risk assessment tool that can help build this comprehensive view of new technology.

10.3 Related Works

Relevant prior work focuses on different ways to reduce vulnerabilities (risk) [5] that employ computational methods using the CVSS methodology, Commercial-Off-The-Shelf (COTS) software, and fuzzy risk analysis to improve control measures or improve system survivability.

The CVSS Risk Estimation Model is a security trade-off decision-support engine for balancing security with a cost. The model's cost-benefit perspective of a security reflects financial and project factors such as budget and time-to-market [7]. However, accurate risk level measurement remains a challenge. The security attributes outlined in the CVSS Risk Estimation Model applies to system service levels [8]. The model relied on experts to define the service levels but does not contain aggregating experience-based information sources.

Chen et al. discussed using CVSS for COTS software systems to measure security investment benefits [3, 4]. They argue that the CVSS does not take the context of the values into account and is misleading. The authors proposed an Analytic Hierarchy Process (AHP) for the stakeholders-values such as the productivity, reputation, and privacy of systems, rather than using the CVSS environmental metric group attributes. Stakeholders have different perceptions of the extent to which a vulnerability might affect them. Both productivity and reputation are subjective and equally hard to estimate, as are the environmental metric group attributes.

Dondo's approach used fuzzy risk analysis for vulnerability prioritization to derive the risk level or risks to a system where the asset value is assumed to be known [6]. The author derives the risk level as $AV \times I \times L$ and applies fuzzy rules that compute impact and likelihood. This approach is like the Houmb et al. Risk Level Estimation Model [8] but differs in the risk level estimate using asset value and safeguards rather than the CVSS temporal and environmental metric group attributes. Asset value is specific to the context and stakeholder, which is not necessarily easy to evaluate.

NIST Special Publication 800-53 presented a proactive and systematic approach to developing a comprehensive safeguard measure for all computing platforms, including general-purpose computing systems [1]. It suggests that the remaining risks represent potential threats to the system. However, the NIST framework lacks contextual knowledge to determine acceptable risks based on budget, time, and resource constraints.

The Cyber Survivability Endorsement (CSE) is the critical foundation for ensuring Cyber Survivability Attributes (CSAs) are considered part of the operational risk trade-space [14]. CSE leverages the NIST 800-53 cybersecurity technical controls but does not define any new cybersecurity requirements. It helps sponsors understand risk, and articulate survivability requirements, with CSAs driving the analysis guidance and acquisition source selection criteria to justify specific cyber mitigations and risk management framework technical controls. CSE provides a holistic approach to determine a system's Cyber Survivability Risk Category (CSRC), and then to assess and manage its Cyber Survivability Risk Posture (CSRP) throughout its lifecycle. CSE determines the acceptable risks given budget, time, and resource constraints based on the most recent activity.

10.4 Methodology

Our method first searches the Department of Homeland Security NVD and groups by technologies, which aggregate as inputs for a distribution estimate of the CVSS factors. Our approach builds upon the Risk Estimation Model of Houmb et al. [7], which uses CVSS information and derives a security

risk level from the CVSS vulnerability information. To this, we add a combination of exploitability and impact estimates. We associate exploitability with an impact that differentiates from past CVSS efforts based on frequency. We implement a BBN that extends the model using Pourret et al. [14] and Jensen [9], extending the topology that uses CVSS-based estimates but also a combination of disparate information sources.

The second element of our methodology uses a Common Vulnerability Scoring System (CVSS) [5]. In our model, CVSS information derives a security risk level from vulnerability information as a combination of exploitability and impact estimates. The NVD selected CVEs applied CVSS metrics serve as prior knowledge for Bayesian applications. The BBN implements a topology that uses CVSS-based estimates as a combination of disparate information sources. The advantage of the BBN is its ability to use whatever additional risk information is available.

10.5 Implementation

Our implementation begins with the NVD search for the CVE incidents by 12 groups of technology security by generation, from the research of Partha Pal and Hamed Okhravi [12, 13]. The NVD files are consolidated using a JSON to CSV converter for a total of 2,691,536 records. The first review retained 123,467 incidents identified by a CVE.

Figure 10.1, Data Framework showed these incidents reduced to 54,580 incidents for complete records with no null states. Lastly, the set reduced to 1355 incidents where the key terms for the vulnerability matched one or more of the

Table 10.1 Key terms result

NVD key terms – results		
Technology	Json files	NVD search
IDS	807	1420
VPN	202	535
Access controls	72	187
PKI	25	166
Cryptography	31	63
Dynamic data		63
Firewall	216	57
Dynamic platform	1	5
Dynamic software		20
Dynamic runtime		3
Boundary controller		2
Trusted computing	1	1
	1355	2522

12 selected security technologies by generation (Table 10.1). Independent subgroups are aggregated separately using the CVSS factors and provide the BBN distribution inputs.

We change the vulnerability states to a CVSS value to facilitate a numeric statistical distribution.

The BBN utilizes both the NVD data distribution and CVSS (version 3.1) metrics for the exploitability and impact scores, and the BBN incorporates their probabilities. We use the BBN to calculate the sensitivity of findings for the exploitability and impact of the existing technologies' vulnerabilities.

10.6 Dataset

The NVD Key Term chart notes a difference between the number of incidents in the keyword searches on the NVD site, and this study's processed JSON NVD Files. Incidents did not include the CVSS V2 columns due to its physical bias and 2020 incomplete year.

Also, we removed items with sample sizes less than 25, which removed the Dynamic (Platform, Software, Runtime & Data) data, and non-selection of the CVSS V2 columns removed the earlier technologies of Boundary Controller and Trusted Computing.

10.7 Data Validations

The Statistical Package for the Social Sciences (SPSS) release 26.0 was used to provide descriptive statistics, bivariate statistics for Means, ANOVA, Correlation, and a linear regression model. The dataset is assumed to have a normal distribution and has the following properties necessary for linear regression [16]:

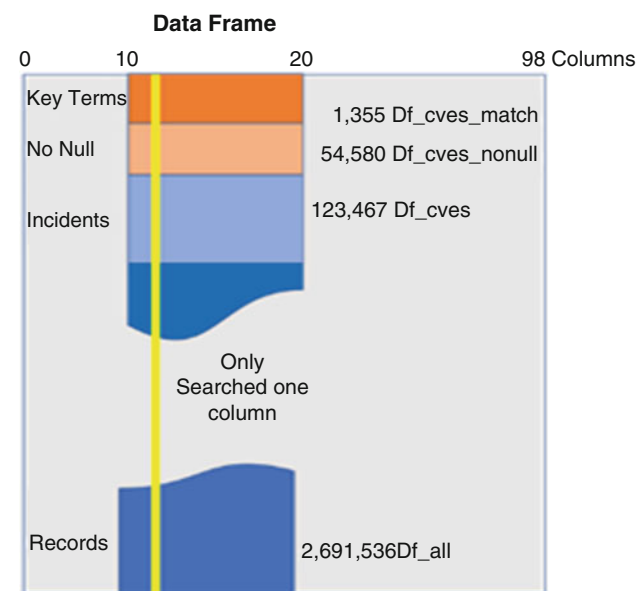


Fig. 10.1 Data Framework

Fig. 10.2 Exploitability scatterplot

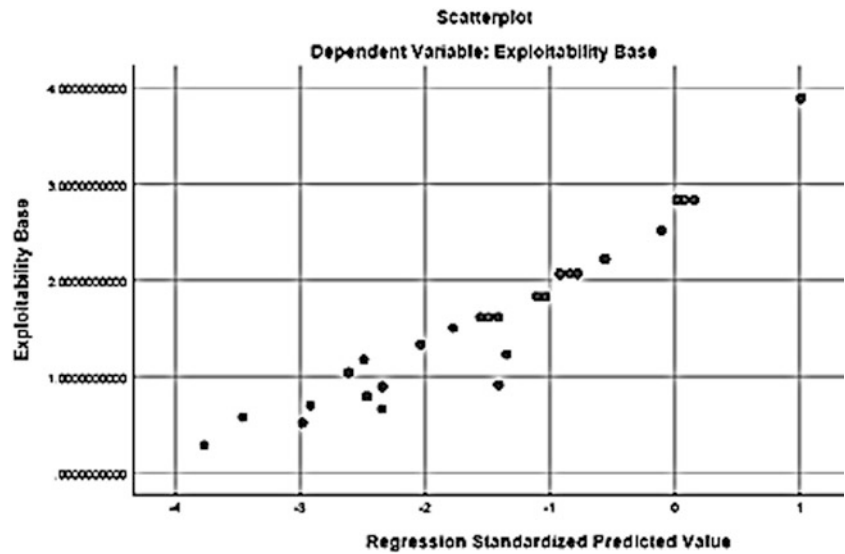
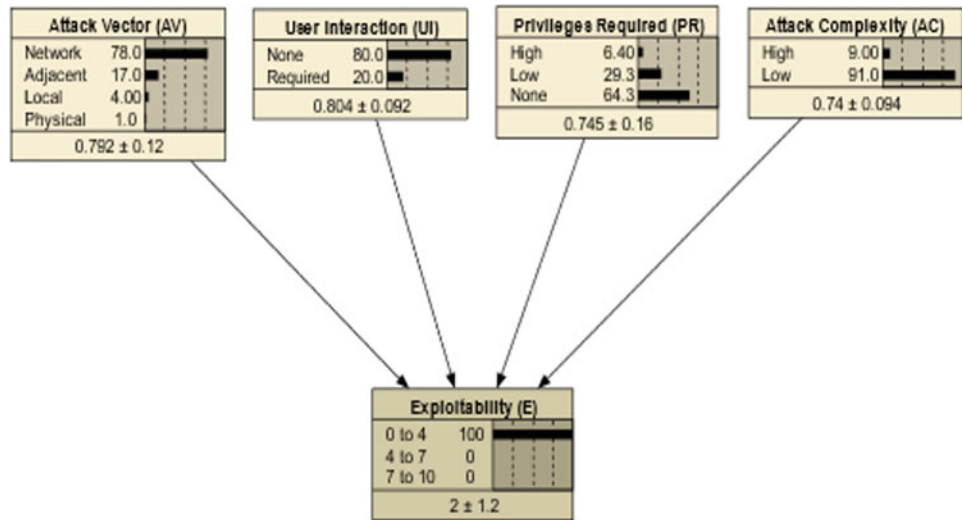


Fig. 10.3 CVSS BBN exploitability



- Linear Relationship
- Homoscedasticity
- No or little Multicollinearity
- No Autocorrelation
- Multivariate Normality

Each property above was tested on the 1355 incidents, where the Linear Relationship and Homoscedasticity assumptions were best tested by scatter plots, shown in Fig. 10.2. An essential check is for outliers since linear regression is sensitive to outlier effects and Homoscedasticity. These are quick, visible checks; the remaining results are part of the discussion section.

10.8 BBN Topology

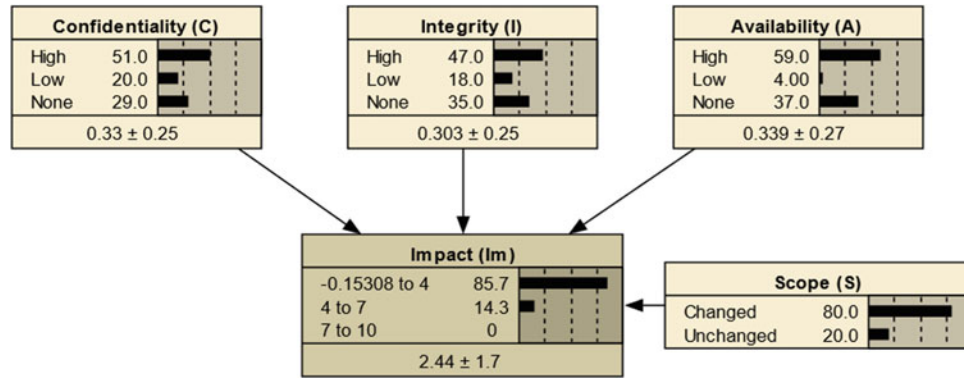
The BBN topology provides a seamless aggregation of CVSS known vulnerabilities that represent new technologies' cy-

bersecurity within a consistent, flexible, and open manner reflective of an organizational context and business domain [9]. Additionally, the BBN provides a mathematically accurate way of assessing different events (or nodes, in this context) effects on each other. Assessments made in either direction can compute the most likely effects given the values of specific causes and determine the most likely causes of observed events.

The BBN topology allows multiple abstraction layers of the input information to derive exploitation, impact, and risk level estimates. The node and their states are derived from FIRST CVSS [5], and their distributions are the NVD reported distributions. The CVSS BBN Exploitability and Impact in Figs. 10.3 and 10.4 represent a proposed technology computation using vulnerability information of existing technology to a CVSS baseline.

The BBN is an interactive model that allows a user to select states that apply defined CVSS equations for the Exploitability or Impact child nodes results. Moreover, the

Fig. 10.4 CVSS BBN impact



Exploitability or Impact node belief bars show each state’s conditional probabilities.

The CVSS Exploitability sub-score (Eq. 10.1) is derived from the Base Exploitability metrics, while the Impact sub-score (Eqs. 10.2 and 10.3) are derived from the Base Impact metrics [5].

Exploitability

$$E (AV, AC, PR, UI) = 8.22 * AV * AC * PR * UI \quad (10.1)$$

Impact

$$Im (I, C, A, S) = S == \text{Unchanged?} \quad (10.2)$$

$$6.42 * (1 - ((1-C) * (1-I) * (1-A)))$$

$$Im (I, C, A, S) = S == \text{Changed?} \quad 7.52$$

$$* ((1 - ((1-C) * (1-I) * (1-A))) - 0.029)$$

$$- 3.25 * (((1 - ((1-C) * (1-I) * (1-A))) - 0.02)$$

$$(10.3)$$

10.9 Sensitivity of Findings

Sensitivity analysis reveals how much a single finding could influence the target node’s beliefs and mean value at each of the other nodes in the net [11]. The analysis reports the minimum and maximum beliefs for each state as 0 and 1, respectively, and the maximum reductions in variance and entropy will be 100%. The expected reduction in the query variable variance has an expected real value, as shown in Tables 10.2 and 10.3 for each variable, where a higher value is better. The mutual information is the expected reduction in the query variable’s entropy due to a finding affecting other variables and a higher value is better. The Variance of beliefs is the expected change (squared) of the query variable’s beliefs, taken over all its states, due to a finding at other variables. A lower value is the least disruptive.

Table 10.2 Exploitability sensitivity of findings

Node	Variance Reduction	Percent	Mutual Info	Percent	Variance of Beliefs
User interaction (UI)	0.008464	100	0.72193	100	0.160000
Attack complexity (AC)	0.008919	100	0.43647	100	0.081900
Attack vector (AV)	0.0135	100	0.96637	100	0.201530
Privileges required (PR)	0.02611	100	1.18238	100	0.284476

Table 10.3 Impact sensitivity of findings

Node	Variance Reduction	Percent	Mutual Info	Percent	Variance of Beliefs
Confidentiality (C)	0.06098	100	1.47772	100	0.3966400
Impact (Im)	0.002243	3.68	0.04567	3.09	0.0052984
Integrity (I)	0.06442	100	1.48736	100	0.4009300
Impact (Im)	0.003144	4.88	0.05730	3.85	0.0073317
Availability (A)	0.0719	100	1.16560	100	0.2828960
Impact (Im)	0.004048	5.63	0.06250	5.36	0.0121648

When the sensitivities are calculated, all findings currently entered in the network will significantly affect the sensitivities. Tables 10.2 and 10.3 below report each varying node, showing how much it can affect the query node using several different sensitivity measures. Tables 10.4 and 10.5 are summary tables that compare the sensitivities for each of the varying nodes, showing the lowest state values (Table 10.4) and the highest state values (Table 10.5).

10.10 Discussion

The NVD files provide distributions of past known vulnerabilities that we use to assess security technologies’ next

generation. The NVD data are separated as independent subgroups and provide the probability distribution inputs for the BBN models' CVSS factors. Previous efforts did not include the exploitability and impact of known vulnerabilities.

Our methodology takes advantage of the publicly available NVD data. Figure 10.2 scatterplots show the data satisfies the linear regression assumptions for a linear relationship. The data histogram confirms the multivariate normality assumption. The assumptions for Multicollinearity assumption confirmed by a Tolerance <0.1 and Variance Inflation Factor < 5 . Condition Index Values 10–30 only indicate mediocre Multicollinearity.

Durbin-Watson's d tests show some Autocorrelation as 0.738 is outside the rule of thumb values of $1.5 < d < 2.5$ for no autocorrelation in the data. However, the Durbin-Watson's d value tests only first-order effects of linear autocorrelation between direct neighbors. The NVD data of the CVE incidents support the necessary data assumptions for linear regression. The substantial Regression in ANOVA of $F(4, 1350)$ is further evident in the R and R^2 of the NVD data.

The BBN sensitivity to findings shows how much findings at other nodes influence a node. Tables 10.4 and 10.5 show the difference in an expected reduction in variance and entropy of the expected real value due to a finding. User Interaction and Confidentiality are more significant nodes. Simultaneously, the variance of node belief provided the expected change (squared) over all its states due to a finding and quantified influences from other nodes' findings.

The BBN topology provides a more seamless aggregation of the CVSS based estimate vulnerabilities than in other

Table 10.4 Lowest state probabilities nodes selected

(a) Exploitability			
Node	State	New finding	All findings
Attack vector (AV)	Physical	1%	
User interaction (UI)	Required	20%	0.2%
Privileges required (PR)	High	6.4%	0.0128%
Attack complexity (AC)	High	9%	0.00115%

Note: With an Exploitability (E) Mean 0.121109, \pm 0.0 Std. Dev., 0.121109 Median, 0.0 Interquartile Range (IQR)

(b) Impact			
Node	State	New finding	All findings
Confidentiality (C)	None	29%	29%
Integrity (I)	None	35%	10.15%
Availability (A)	None	37%	3.7555%

Note: Scope (S) Changed State Selected with Probability of new finding = 20%, of all findings = 0.7511%. With an Impact (Im) Mean 1.92346, \pm 1.19889 Std. Dev., 1.92346 Median, 2.07654 IQR
Scope (S) Unchanged State Selected with Probability of findings = 3.0044%. With an Impact (Im) Mean - 0.15308, \pm 0.0 Std. Dev., -0.15308 Median, 0.0 IQR

Table 10.5 Highest state probabilities nodes selected

(a) Exploitability			
Node	State	New finding	All findings
Attack vector (AV)	Network	78%	
User interaction (UI)	None	80%	62.4%
Privileges required (PR)	None	64.3%	40.123%
Attack complexity (AC)	Low	91%	36.5121%

Note: With an Exploitability (E) Mean 3.88704, \pm 0.0 Std. Dev., 3.88704 Median, 0.0 IQR

(b) Impact			
Node	State	New finding	All findings
Confidentiality (C)	High	51%	51%
Integrity (I)	High	47%	23.97%
Availability (A)	High	59%	14.1423%

Note: Scope (S) Changed State Selected with Probability of new finding = 20%, of all findings = 2.82846%

With an Impact (Im) Mean 3.75318, \pm 0.0 Std. Dev., 3.75318 Median, 0.0 IQR

Scope (S) Unchanged State Selected with Probability of findings = 3.0044%. With an Impact (Im) Mean 5.87312, \pm 0.0 Std. Dev., 5.87312 Median, 0.0 IQR

work, and the NVD CVE data satisfies the Bayesian assumption of prior knowledge that was not evident in previous works.

10.11 Conclusion

This methodology presents a measurable security risk level estimation for the next generation of security technologies that applies the CVSS for any computerized system's critical factor security. Table 10.2 shows that the Privilege Required has the most influence in the variance reduction with the highest mutual information and variance of belief. Furthermore, to a varying degree, Attack Complexity and Attack Vector are less influential. User Interaction is the least important with the following evaluation criteria. The results in Table 10.3 show that Integrity follows Availability and then Confidentiality for the most influence in the variance reduction with the highest impact mutual information, percent, and variance of belief. The BBN sensitivity to findings shows the importance of the factors that affect the hybrid security risk assessment model for new technologies.

The methodology in this work takes advantage of the NVD publicly available data for the distribution used in the subsequent CVSS analysis and BBN construction. NVD CVE incidents provide the prior knowledge that satisfies a fundamental assumption for the BBN, where previous efforts assumed normal distribution. The flexibility provided by the BBN topology allows other sources to combine with the

CVSS information and various levels of abstraction in the input information. The sensitivity analysis determines the most significant nodes in forming the beliefs of the key nodes. That changes as findings arrive, so it may need to be recomputed at each stage.

In future work, we will use the CVSS distributions of known vulnerabilities to provide prior knowledge for CWSS-based estimates in a BBN risk model for new technologies.

References

1. T. Allen, NIST Cybersecurity for IoT Program—Publications. NIST. <https://www.nist.gov/itl/applied-cybersecurity/nist-cybersecurity-iot-program/publications> (2019, September 4)
2. “CAPEC – About CAPEC.” 2019. <https://capec.mitre.org/about/index.html> (February 6, 2020)
3. Y. Chen, Stakeholder value driven threat modeling for off the shelf based systems. 29th International Conference on Software Engineering (ICSE’07 Companion), pp. 91–92. <https://doi.org/10.1109/ICSECOMPANION.2007.69>, 21 citations (2007)
4. Y. Chen, B. Boehm, L. Sheppard, Measuring security investment benefit for off the shelf software systems – a stakeholder value driven approach. Online Proceedings of Sixth Workshop on the Economics of Information Security (WEIS 2007) (2007)
5. “CVSS v3.1 Specification Document.” FIRST — Forum of Incident Response and Security Teams. <https://www.first.org/cvss/specification-document> (October 4, 2020)
6. M.G. Dondo, A vulnerability prioritization system using a fuzzy risk analysis approach. Proceedings of The Ifip Tc 11 23rd International Information Security Conference, pp. 525–540. https://doi.org/10.1007/978-0-387-09699-5_34, 22 citations (2008)
7. S.H. Houmb, V.N.L. Franqueira, Estimating ToE risk level using CVSS. Proceedings of the Fourth International Conference on Availability, Reliability and Security (ARES 2009 – The International Dependability Conference), pp. 718–725. <https://doi.org/10.1109/ARES.2009.151>, 52 citations (2009)
8. S.H. Houmb, V.N.L. Franqueira, E.A. Engum, Quantifying security risk level from CVSS estimates of frequency and impact. *J. Syst. Softw.* **83**(9), 1622–1634 (2010) <http://www.sciencedirect.com/science/article/pii/S0164121209002155> (October 20, 2019)
9. Introduction to Bayesian Networks—Finn V. Jensen—Google Books, https://books.google.com/books?id=g8hlQgAACAAJ&dq=Jensen,+F.,+1996.+An+Introduction+to+Bayesian+Network.+UCL+Press,+University+College+London.&hl=en&newbks=1&newbks_redir=0&sa=X&ved=2ahUKEwjX2q_Z7bvnAhXqdd8KHa9oAfMQ6AEwA3oECAIQAg, 4621 citations (1997)
10. “NVD – Data Feeds.”. <https://nvd.nist.gov/vuln/data-feeds> (February 6, 2020)
11. Norsys, Welcome to Netica’s Help System, <https://www.norsys.com/WebHelp/NETICA.htm>, 2020
12. H. Okhravi, M.A. Rabe, T.J. Mayberry, W.G. Leonard, T.R. Hobson, D. Bigelow, W.W. Streilein, Survey of cyber moving target techniques. DTIC Document. <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA591804>, 177 citations (2013)
13. P. Pal, R. Schantz, M. Atighetchi, J. Loyall, F. Webber, What next in intrusion tolerance. BBN Technologies, Cambridge. <http://wraits09.di.fc.ul.pt/wraits09paperParthaPal.pdf>, 8 citations (2009)
14. J. Petty, Cybersecurity Test and Evaluation Guidebook 2.0, https://daytonaero.com/wp-content/uploads/DOD_Cybersecurity-Test-Evaluation-Guidebook-ver-2.0_25-APR-2018.pdf (2018)
15. O. Pourret, P. Naïm, B. Marcot, *Bayesian Networks: A Practical Guide to Applications*. John Wiley & Sons, 443 citations, <https://vbn.aau.dk/en/publications/an-introduction-to-bayesian-networks> (2008)
16. A. Santos-Caballero, Why is it important to examine the assumption of linearity when using regression? ResearchGate. https://www.researchgate.net/post/Why_is_it_important_to_examine_the_assumption_of_linearity_when_using_Regression (2018, February 1)

Enriching Financial Software Requirements Concerning Privacy and Security Aspects: A Semiotics Based Approach

Leonardo Manoel Mendes, Ferruccio de Franco Rosa, and Rodrigo Bonacin

Abstract

Enriching software requirements with key security and privacy features requires professionals to have knowledge of requirements elicitation techniques, based on systematic processes and methods. We propose the Software Requirements Analysis Method for Improvement of Privacy and Security (SRAM-PS), which is based on concepts and techniques from Organizational Semiotics and on the analysis of information security and data privacy standards. SRAM-PS is a 7-steps systematic approach where an input set of software requirements is analyzed, processed, and then enriched with new security and privacy requirements. A case study with 4 experts was carried out, where SRAM-PS is used in a real world scenario: a bank sends a financial transaction receipt containing the customer's personal data over the Internet. SRAM-PS is aimed at researchers and engineers who analyze and specify software requirements and need to systematize their methods and techniques.

Keywords

Organizational semiotics · Requirements engineering · Privacy · Information security · Financial software

L. M. Mendes
University of Campo Limpo Paulista (UNIFACCAMP), Campo Limpo Paulista, SP, Brazil

F. de Franco Rosa · R. Bonacin (✉)
University of Campo Limpo Paulista (UNIFACCAMP), Campo Limpo Paulista, SP, Brazil

Renato Archer Information Technology Center (CTI), Campinas, SP, Brazil
e-mail: rodrigo.bonacin@cti.gov.br

11.1 Introduction

Specifying requirements of critical systems properly addressing privacy and security issues is a need for different stakeholders in software development. The need for privacy and security is imperative, however, there are gaps in the analysis and specification of different dimensions, from the views of users and organizations.

Currently, when eliciting software requirements, we must consider principles contained in international data protection and privacy regulations, e.g., European General Data Protection Regulation (GDPR),¹ California Consumer Privacy Act (CCPA),² Brazilian Data Protection General Law (LGPD),³ in addition to knowledge sources and information security standards (e.g., [1]).

New methods and techniques aimed at assisting software engineers in identifying privacy and security requirements are required. Although there are techniques for eliciting software requirements, there are gaps in techniques for enriching requirements with respect to important aspects of privacy and security, as these involve technical, legal, organizational, social issues, and other subjective factors that are difficult to systematize.

We identified works with a semiotic approach in the organizational context to facilitate communication between Stakeholders [2], as well as works that address the issue of privacy in eliciting requirements [3]. However, there is a lack of methods to improve privacy and security for software requirements, particularly considering human aspects on financial systems (our focus).

¹<https://gdpr-info.eu/>

²<https://oag.ca.gov/privacy/ccpa>

³ http://www.planalto.gov.br/ccivil/_03/_ato2015/-2018/2018/lei/113709.htm

We propose the Software Requirements Analysis Method for Improvement of Privacy and Security (SRAM-PS), which comes from concepts and techniques of Organizational Semiotics and brings a sociotechnical view to the elicitation process. Building the proposed method was a multidisciplinary research challenge, as it involved skills in requirements engineering, semiotics, data privacy and information security. The combination of concepts and techniques from various areas aims to mitigate important gaps, such as the lack of understanding of the human aspects of privacy.

We start from the following research hypothesis: a method based on organizational semiotics, agile requirements analysis concepts and knowledge sources in security and privacy is able to better assist in the enrichment of financial software security requirements. The main contributions are: (i) a method based on semiotics to improve aspects of privacy and security during the stage of analysis, elicitation and specifications of financial software requirements; and (ii) a case study, containing a scenario of application of the method.

The remaining of this article is organized as follows: In Sect. 11.2, we present a background and the related work; in Sect. 11.3, we present the SRAM-PS method; in Sect. 11.4, we present a case study of applying SRAM-PS; and in Sect. 11.5, we present the conclusion and final remarks.

11.2 Background and Related Work

Firstly, we present an introduction on the requirements elicitation and on secure development, then we address the organizational semiotics and, finally, we presents the related work.

11.2.1 Requirements Elicitation and Secure Systems Development

According to [4], several requirements analysis and specification process is composed of 4 phases: (1) requirements discovery, (2) classification and organization of requirements, (3) prioritizing and negotiating of requirements, and (4) requirements specification. During requirements elicitation, software engineers use techniques to discover functional and non-functional requirements of the system. For example, to elicit security requirements for software systems projects in the organization, we need to know vulnerabilities that can lead to cyberattacks or confidential data leaks [5].

There are also proposals for secure software development processes, e.g., the Microsoft SDL [6], which is a set of practices that support the specification of requirements aimed at compliance with security standards. However, the development of secure software requires the adoption of systematic and systemic (comprehensive) methods.

11.2.2 Organizational Semiotics

Semiotics is the science study meaning and of all types of signs. Organizational Semiotics considers organizations as information systems in which information is created, processed, distributed, stored and used, developing a semiotic theory [7].

MEASUR (Methods for Eliciting, Analyzing and Specifying User Requirements) [7, 8] aims to constitute a set of methods and techniques for eliciting, analyzing and specifying requirements based on the concepts of semiotics. As part of MEASUR a subset called the Problem Articulation Method (PAM) is used to clarify problems when they are still vague. Artifacts and techniques of PAM described below were used in this work.

The *Stakeholders Diagram (Semiotic Onion)* is used to analyze stakeholders according to their degree of relation with the system. In this diagram, stakeholders are classified into layers, where the more internal ones mean closer relation with the system. Requirements engineers, customers and users should agree and define the focus of the analysis. For example, in the context of this work, a structured brainstorming can be carried out with a focus on stakeholders related to security and privacy requirements. The diagram consists of 5 layers, namely: Operation, Contribution, Source, Market and Community.

The *Evaluation Framework* elicits problems and possible solutions for the system. For this purpose, possible problems that exist or that could arise during the system development are identified. In sequence, we need to identify and propose possible solutions to mitigate the problems. The framework is filled out in a brainstorming, where questions and problems are assigned to each stakeholder of each layer of the stakeholder diagram. Subsequently, possible solutions to the problems are defined.

The *Semiotic Framework (or Semiotic Ladder)* aims to assist software engineers in eliciting and specifying requirements in different levels of semiotics. It is divided into two parts: (i) Functions of the Human Information System, which include the social, pragmatic and semantic worlds, and (ii) Information Technology Platform, which includes the syntactic, empirical and physical elements.

11.2.3 Related Work

This subsection focuses on the intersection areas between semiotics, information security, data privacy, and requirements. Our review methodology is based in [9, 10]. We used the following search string in scientific databases: *semiotics*

AND requirements AND (trust OR security OR privacy). Twenty-three papers were analyzed, and according to their contributions, 5 selected papers are described.

Organizational semiotics is applied in [3] for extending the technology acceptance model for proximity mobile payment. The semiotic ladder is used to assess acceptance in human information system and technology platform to identify the requirements for adopting mobile payment. In this work, only one technique (semiotic ladder) is used. A framework for managing conflicts between non-functional security and privacy requirements is proposed in [11] aiming at reducing the risk impact on the software development project. A framework for requirements modeling, which considers aspects of security, privacy and trust in the health area is proposed in [12]. An approach that uses semiotic inspection techniques for the requirements eliciting activity is proposed in [13], in which mind maps are created to characterize a certain functional requirement. An experimental protocol is also used to verify the meta-communication between the proposed artifact and the artifact that was generated. An approach for health systems development based on organizational semiotics is proposed in [14]. User requirements are elicited and analyzed by means of ontology diagrams and standard modeling in order to standardize requirements of the health domain. Other works (e.g., [2]) use semiotics as an instrument for the identification and modeling of requirements. However, our work differs from the others since it is the only one to use organizational semiotics to enrich software requirements considering critical aspects of privacy and security.

11.3 Software Requirements Analysis Method for Improvement of Privacy and Security (SRAM-PS)

As shown in Fig. 11.1, SRAM-PS is a 7-step method that ranges from receiving a set of system requirements (Fig. 11.1 – A1) to improving these requirements with an additional set of privacy and security requirements (Fig. 11.1 – A8). The steps should be taken/adapted by an Information Security Engineer (ISE), when specifying requirements. Each step contains a set of tasks to be performed to generate artifacts. The artifacts can vary according to the needs, characteristics and methodologies to be used for the software under development (e.g., user stories, acceptance criteria or use cases). The SRAM-PS steps are detailed as follows.

Step 1: Identify Knowledge Sources of Security and Privacy Requirements (KSSPR) - We identify the KSSPRs that should be used in the process of eliciting or improving requirements of the system under development. Examples of KSSPR: PCI-DSS [15], ISO/IEC 27002 [1], OWASP

Testing Guide,⁴ among other test sets or regulations related to the international financial system. The use of each KSSPR depends on the privacy and security needs.

Step 2: Identify Stakeholders - We identify in a brainstorm stakeholders who can influence or be influenced directly or indirectly by the privacy and security requirements related to the system.

Step 3: Analyze Stakeholders - Analyze the Stakeholders candidates using the semiotic onion layers, namely: (i) Community: outermost layer, in which should be listed, e.g., observers who exercise indirect influence or are influenced by the system's privacy and security; (ii) Market: it includes stakeholders from the market, e.g., partners who have interests related to the system's privacy and security; (iii) Source: it includes system sources, such as customers and suppliers, who have interests related to the system's privacy and security; (iv) Contribution: it includes stakeholders who directly contribute to the system, such as users of the system, system administrators, among others who directly influence or are influenced by aspects related to the system's privacy and security; and (v) Operation: it is the innermost layer, where the system is, internal stakeholders related to the system's privacy and security are part of this layer.

Step 4: List Stakeholders' Interests – List the privacy and security interests and needs of the stakeholders that were analyzed in the last step. We use strategies to reconcile discordant interests, such as workshops, interviews and document analysis. The security and privacy needs must be presented in a clear and objective manner

Step 5: Analyze Problems and Ideas – We build the evaluation framework to identify and analyze existing privacy and security problems, as well as to propose solutions. The existing problems related to each stakeholder must be identified and, based on these problems, questions related to the problems must be formulated so that the stakeholders can propose solutions. The evaluation framework is organized in levels according to the semiotic onion, each problem or idea related to privacy and security is linked to semiotic onion' stakeholders.

Step 6: Evaluate Requirements – We develop the semiotic framework, in which the 6 levels are filled:

- *Describe Requirements of the Social World*, such as legal aspects and society's perception of privacy and information security aspects;
- *Describe Requirements Related to Pragmatics*, such as the user's intentions and commitments concerning privacy and information security;
- *Describe Requirements Related to Semantics*, such as whether users understand the language used to describe privacy and information security issues and terms;

⁴<https://owasp.org/www-project-web-security-testing-guide/>

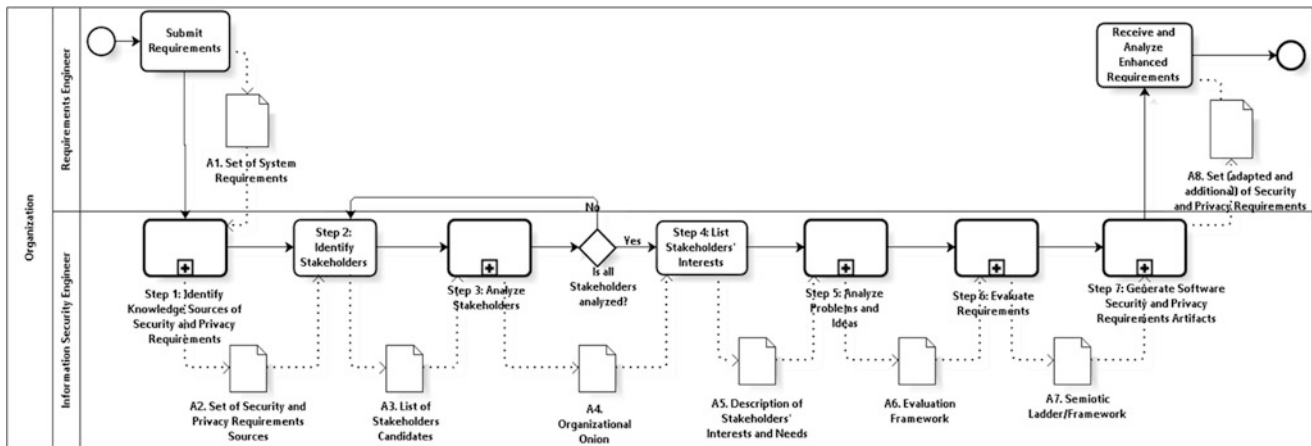


Fig. 11.1 Process for running MARS

- *Describe Requirements Related to Syntax*, such as programming languages, protocols and data models used and their relationship with privacy and information security;
- *Describe Requirements Related to Empirical Aspects*, such as communication bandwidth, speed and processing capacity and their relation to privacy and information security;
- *Describe Requirements Related to Physical Aspects*, such as computers, processors, physical networks and their relation to privacy and information security aspects.

Step 7: Generate Software Security and Privacy Requirements Artifacts - This step aims to generate output according to the required format, characteristics, needs and methodology adopted by the project. This includes creating artifacts such as: *User Stories*, for agile projects, *Acceptance Criteria*, for hybrid approaches, and *Use Cases*, for UML based projects. In all cases, the artifacts must specify the needs, requirements, ideas, problems and solutions listed in the previous steps.

11.4 Case Study of Applying SRAM-PS

We present the case study of applying SRAM-PS. Subsection 11.4.1 describes the evaluation method and Subsection 11.4.2 the results and discussions.

11.4.1 Objectives, Participants and Method

The objective is to verify the feasibility of SRAM-PS in real-world. A workshop was carried out with participation of experts in requirements and information security who work in the financial field. These specialists analyzed the use of SRAM-PS in the following scenario: sharing receipts of

bank transactions carried out in internet banking and mobile banking. Such receipts contain sensitive information that could be shared without any users' verification, leading to risks related to privacy and security. Thus, there is a need for a rigorous analysis of the personal data and requirements.

Four experts participated in the workshop, 2 from the information security and 2 from the requirements engineering. All of them work in financial companies: 3 from São Paulo, and 1 from Miami. Profiles of experts: 1 information security engineer of a bank; 1 commercial coordinator of an information security company; 1 business analyst of credit card flag; and 1 information security engineer of a loan company. Participants have 10, 5, 4 and 5 years of experience in their roles, respectively.

The Workshop procedure: (1) the researcher made a presentation of the SRAM-PS, detailing the activities and giving examples of artifacts produced in each step; (2) the scenario was presented highlighting aspects of security and privacy; (3) the experts executed the SRAM-PS to enrich the requirements; and, (4) the experts answered quantitative and qualitative questions. The method was executed individually, without communication between the participants, during 3 h:30 m.

11.4.2 Results and Discussion

In step 1, the experts suggested sources of requirements related to the financial system (e.g., PCI-DSS [15]), vulnerability databases and sources of security and privacy requirements (e.g., OWASP Testing Guide). For a deeper understanding they suggested to use vulnerability risk prioritization technologies, such as Kenna Security⁵ and Tenable.io;⁶

⁵<https://www.kennasecurity.com/>

⁶<https://www.tenable.com/>

they could address vulnerabilities to prioritize the critical ones. One of the experts considered the sources sufficient for the scenario, but also recommended considering other frameworks (e.g., NIST,⁷ COBIT⁸). Another expert suggested GDPR, CCPA, LGPD, and ISO/IEC 27002. *Results:* 8 knowledge sources were listed.

In step 2, the experts stated that the activity of identifying stakeholders was effortless. They pointed out the difficulties with the availability of stakeholders. *Results:* 8 stakeholders were listed.

In step 3, experts considered that the organizational onion provides a holistic view of the roles and responsibilities of all stakeholders. They also highlighted the advantage of clearly defining the perceptions of each stakeholder in each layer of the onion. *Results:* The roles of 8 stakeholder profiles were defined and analyzed.

In step 4, the experts highlighted difficulties in imagining and meeting the needs and desires of each stakeholder. Three experts agreed that stakeholders could have difficulty expressing their needs, as they might not understand requirements, aspects of security technology, problems with cost and time, etc. Additionally, the lack of a culture of security and privacy among stakeholders was pointed out as a potential problem. *Results:* 18 privacy and security needs have been identified.

In step 5, experts pointed out the advantages and disadvantages of the evaluation framework. As advantages, experts pointed that it is useful for defining and prioritizing requirements and pre-existing vulnerabilities. The visibility and organization that the evaluation framework provides for the requirements was also pointed out, allowing us to analyze and discuss the impact of each requirement and to propose improvements. As disadvantages, the experts identified that: (i) it can be time consuming, mainly by listing solutions that cannot be applied; (ii) the number of ideas and solutions can be very large, increasing thus the size of the project. *Results:* 5 problems were defined and prioritized; 4 pre-existing vulnerabilities identified.

In step 6, they pointed out the semiotic framework is suitable to dynamically manage security and privacy requirements. It allows us to engage stakeholders and provide a personalized view of the organization's processes. An expert pointed out the semiotic framework can be useful for analyzing the company's security maturity level. As a disadvantage, it was pointed out the use of the semiotic ladder takes time and is bureaucratic, making the process less agile. *Results:* 10 potential requirements have been defined and prioritized.

In step 7, experts reported difficulties in choosing the appropriate artifact for the methodological approach of their respective companies. However, they considered artifacts ca-

pable of representing the improved requirements. An expert suggested that the term "artifact" should be changed, since it is a common term used in the field of computer forensics. *Results:* 2 initial requirements have been improved; 8 new privacy and security requirements have been proposed.

In general, SRAM-PS presented advantages, such as providing engineers with a systematic process and a broad view of what needs to be done to ensure security and privacy. According to the experts, it is possible to avoid rework and mitigate risks in improving requirements. However, to involve all stakeholders to participate actively in all steps of the process is a hard task; we consider this a limiting factor in the application of the proposed method in a real scenario, and therefore, it demands research to improve the method.

11.5 Conclusion and Final Remarks

We presented the Software Requirements Analysis Method for Improvement of Privacy and Security (SRAM-PS), which is based on concepts and techniques of organizational semiotics and on the analysis of information security and data privacy standards. Four experts applied our 7-steps systematic approach in a real world scenario: a bank sends a financial transaction receipt containing the customer's personal data over the Internet. An input set of software requirements was analyzed, processed, and then enriched with new security and privacy requirements. Our proposal is aimed at researchers and engineers who analyze and specify software requirements and need to systematize their methods and techniques.

References

1. ISO/IEC, *Information Technology Security Techniques Code of Practice for Information Security Controls*, International Organization for Standardization, Geneva, CH, Standard, Mar. 2013
2. J.C.D. Reis, A.C.D. Santos, E.F. Duarte, F.M. Gonçalves, B.B.N. de França, R. Bonacin, M.C.C. Baranauskas, Articulating socially aware design artifacts and user stories in the conception of the opendesign platform, in *Proc. of the 22nd International Conference on Enterprise Information Systems – Vol 2*, SciTePress, 2020, pp. 523–532
3. Y.C. Pan, A. Jacobs, C. Tan, S. Askool, Extending technology acceptance model for proximity mobile payment via organisational semiotics, in *Digitalisation, Innovation, and Transformation*, ed. by K. Liu, K. Nakata, W. Li, C. Baranauskas, (Springer International Publishing, Cham, 2018), pp. 43–52
4. I. Sommerville, *Software Engineering*, 10th edn. (Pearson Education Limited, Harlow, UK, 2016)
5. K. Qian, R.M. Parizi, D. Lo, OWASP risk analysis driven security requirements specification for secure android mobile software development, *DSC 2018 – IEEE Conference on Dependable and Secure Computing*, pp. 4–5, 2019

⁷<https://www.nist.gov/cyberframework>

⁸<https://www.isaca.org/bookstore/cobit-5/wcb5b>

6. M. Howard, S. Lipner, *The Security Development Lifecycle: SDL, a Process for Developing Demonstrably More Secure Software*, ser. *Best practices* (Microsoft Press, Redmond, WA, USA, 2006)
7. K. Liu, W. Li, *Organisational Semiotics for Business Informatics* (Routledge, Abingdon, 2014)
8. R. Stamper, *Information in Business and Administrative Systems*, ser. *A Halsted Press Book* (Wiley, New York, NY, USA, 1973)
9. R.R. de Mendonça., F.F. Rosa, A.C.T. Costa, R. Bonacin, M. Jino, OntoCexp: a proposal for conceptual formalization of criminal expressions. In: *16th International Conference on Information Technology-New Generations (ITNG)*, 2019, vol 800. Springer, Cham
10. B. Kitchenham, Procedures for performing systematic reviews, Keele University, Keele, UK, vol. 33, no. 2004, pp. 1–26 (2004)
11. D. Alkubaisy, A framework managing conflicts between security and privacy requirements, in *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, 2017, pp. 427–432
12. R.S. Tejas, S.V. Patel, Security, privacy and trust oriented requirements modeling for examination system, in *Nirma University International Conference on Engineering (NUiCONE)*, 2012, pp. 1–6
13. S.F. de Oliveira, P.V. Martinez, J.A. Fabri, A. L’Erario, A. S. Duarte, J. A. Goncalves, Proposal for semiotics inspection method application in coming artifacts requirements survey activity, in *11th Iberian Conference on Information Systems and Technologies (CISTI)*, 2016, pp. 1–7
14. Y. Hongqiao, L. Weizi, Modeling requirement driven architecture of adaptive healthcare system based on semiotics, in *2009 International Forum on Information Technology and Applications*, vol. 2, 2009, pp. 723–727
15. PCI, Payment Card Industry (PCI) Data Security Standard (DSS) Version 3.2.1, *PCI Security Standards Council*, Wakefield, MA USA, Standard, May 2018

Efficient Design of Underwater Acoustic Sensor Networks Communication for Delay Sensitive Applications over Multi-hop

12

Ahmed Al Guqhaiman, Oluwatobi Akanbi, Amer Aljaedi,
and C. Edward Chow

Abstract

Underwater Acoustic Sensor Networks (UASNs) play a critical role in the remote monitoring of a wide range of time-sensitive underwater applications, such as in the oil/gas pipeline to avoid oil spills. In this type of application, the transmission of collected information to the onshore infrastructure within a period of time is critical. Despite the advantages of UASNs over the limitations of Terrestrial Wireless Sensor Networks (TWSNs), the applicability of UASNs in different use-cases requires further investigation. In this paper, we investigate different MAC protocols and study the impact of non-environmental factors that may degrade performance. We simulate different MAC protocol approaches based on available underwater commercial modems to find the most efficient MAC protocol approach for the oil/gas industry based on core performance metrics. Our extensive simulation results show that the contention-based random access approach is the most suitable for time-sensitive application where the Network Size (NS) followed by Network Load (NL),

Data Rate (DR), and Packet Size (PS), respectively have the strongest impact on delay.

Keywords

Media access control protocols · Quality of service · Underwater acoustic sensor networks

12.1 Introduction

Monitoring of an underwater environment is very critical to keep track of the current status of underwater resources [1]. To efficiently monitor underwater resources, it is required to have underwater sensors to form an Underwater Wireless Sensor Network (UWSN). Commonly, underwater sensors utilize acoustic waves to transmit collected data to the sink at the surface level. Underwater applications are categorized as shallow water, deep-water, and ultra-deep-water. The depth of shallow water can be up to 125 m, while deep-water and ultra-deep water range from 125 to 1,500 m and 1,500 to 10,000 m, respectively [2, 3]. Different applications have different objectives and different Quality of Service (QoS) requirements, which require monitoring different depths.

Underwater Acoustic Sensor Networks (UASNs) have a large number of underwater applications, such as environmental monitoring, disaster prevention, pollution monitoring, military activities, and oil/gas spills detection [1, 4]. Typically, UASNs consist of a wide range of sensors that are distributed strategically to monitor a specific area [5]. Different applications have different requirements; thus, the number of underwater sensors depends on the required coverage area and QoS of an application.

A. Al Guqhaiman (✉)
Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, CO, USA

Department of Computer Networks and Communications, College of Computer Sciences and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia
e-mail: aalguqha@uccs.edu

O. Akanbi · C. Edward Chow
Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, CO, USA
e-mail: oakanbi@uccs.edu; cchow@uccs.edu

A. Aljaedi
College of Computing and Information Technology, University of Tabuk, Tabuk, Saudi Arabia
e-mail: aaljaedi@ut.edu.sa

Monitoring the oil/gas pipeline, which is a time-sensitive application requires minimum delay to receive collected data. To prevent man-made or natural disasters, the QoS of this type of application must be met. In order to meet these requirements, studying the MAC protocol with different underwater commercial modems is critical, as the capabilities of the modem affects the network performance. Furthermore, to design the most efficient communication for UASNs, the impact of Data Rate (DR), Network Size (NS), Packet Size (PS), and Network Load (NL) on the network performance metrics must be well defined. Thus, analyzing the relationship between these factors to the network performance metrics is needed to meet the QoS. In comparison to Terrestrial Wireless Sensor Networks (TWSNs), which utilizes the radio waves, UASNs deploy acoustic waves due to the high attenuation of radio waves in water for packet transmission, which makes meeting the QoS more difficult [6].

UWSNs can transmit collected data between sensors using acoustic, light, magnetic induction, or radio waves. Even though light [7, 8], magnetic induction [9, 10], and radio waves can offer higher bandwidth, they have some limitations. Therefore, the acoustic wave is considered the most reliable to transmit data between underwater sensors in such a severe environment [9]. However, it is still a challenging issue to communicate between sensors using a reliable or high DR due to the characteristics of acoustic channels and limited bandwidth. In addition, UWSNs operate in half-duplex mode where an underwater node can only send or receive packets, but not both at the same time. In this mode, all underwater sensors share a common channel. Therefore, there is a high chance that multiple underwater nodes will interfere with the communication between each other. Hence, the Media Access Control (MAC) protocol plays an important role by controlling which node is allowed to transmit packets, and at what time, to minimize interference between intended parties. To design a reliable and efficient MAC protocol for this type of application, it is also essential to determine which MAC protocol approach is more efficient for monitoring the oil/gas pipeline application in a shallow water environment. Furthermore, it is critical to highlight the factors that have strong impact on the network performance.

Collisions between packets transmitted by multiple nodes are one of the critical challenges that cause communication failure and affect network performance. In multi-hop networks, the chance of collisions of transmitted packets is higher compared to single-hop networks in most MAC protocol approaches. As a consequence, a higher collision rate can degrade the network performance. A collision can occur due to hidden-terminal, exposed terminal, spatio-temporal uncertainty, and near-far problems [11, 12]. Whether using a single-channel or multiple-channel, collisions can occur to control and data packets. To the best of our knowledge, this research is the first attempt to comprehensively study the ef-

fect of MAC protocol approaches for the oil/gas industry over multi-hop networks using underwater commercial modems. Furthermore, no existing paper has analyzed the impact of non-environmental factors to the network performance over multi-hop.

The rest of this paper is organized as follows. In Sect. 12.2, we briefly review the challenges and issues of UASN communication, network architecture, and MAC protocol approaches. Section 12.3 presents some related works. In Sect. 12.4, we discuss the factors that impact the performance metrics and highlight the characteristics of underwater commercial modems. Section 12.5 discusses the performance evaluation and compares the simulation results against those of Aloha, BroadcastMAC, RMAC, and TMAC MAC protocols. Section 12.6 concludes this paper.

12.2 Background Information

An efficiently designed MAC protocol for UASNs must consider several factors that play a pivotal role in enhancing underwater communication. Some factors depend on one another and can impact network performance. Therefore, developers must be aware of the tradeoffs between various factors that affect network performance. Akyildiz et al. [6] presented many crucial factors that must be considered while designing MAC protocols, including network topology (ad-hoc, cluster), NS, DR, hop length (single-hop, multiple-hop), and water application (shallow water, deep-water, and ultra-deep-water).

Much of the existing research in UASNs utilizes the communication between sensors and sink at the surface level in a single-hop approach. Recent studies place an emphasis on multi-hop networks to increase the efficiency and coverage area of the network. In multi-hop networks, using relays to forward packets to reach a destination helps to save energy. Akyildiz et al. [6] reveal that transmitting packets directly to the surface node consumes higher energy than using relays to forward the packets. Therefore, in a multi-hop network, it is more energy efficient to transmit packets through relays compared to direct transmission to the destination.

Generally, the network topology in UASNs can be classified into centralized, distributed, and clustered [13, 14]. The centralized topology can refer to having several nodes, where one node is responsible for managing communication within a network and between different networks. This topology can be formed in star and tree topologies. A star topology can simplify network management over a short-range of communication, but suffers from single-point-of-failure. If the central node fails, the whole network becomes unavailable. In contrast, tree topology resolves the issue of single-point-of-failure by having multiple star topologies at different levels.

Therefore, if a single central node fails in a star topology, only that star network becomes unavailable.

In distributed topology, nodes can directly communicate with each other without a central node. We can further classify this topology into a single-hop network and multi-hop network. In single-hop, all nodes are one-hop away from each other and can cover a small area. In this approach, every node is aware of any ongoing transmission in the network hence avoiding the hidden-terminal and exposed-terminal problems. Conversely, in the multi-hop network, the distance between nodes is more than one-hop and can cover large areas. Therefore, packets must be relayed to reach the destination. However, the chance of collisions is higher compared to the single-hop network. The cluster topology integrates the centralized and distributed topologies by forming a cluster from several nodes. In order to cover a large area in cluster topology, packets must pass through several clusters to reach the destination. Within each cluster, the Cluster Head (CH) is responsible for coordinating communication between cluster members and different clusters. Each cluster can be in star or tree topologies [12].

In practice, the network topology of UASNs can be either ad-hoc based or cluster-based. Ad-hoc based protocols are more appropriate for real-time applications, whereas cluster-based protocols are more appropriate for applications that accept increased delay in exchange for decreased energy consumption and fewer collisions. Furthermore, MAC protocols based on cluster-based approach use Time Division Multiple Access (TDMA) to assign a slot for each node. Each node can only transmit its packets at the beginning of the time slot. Although the cluster-based approach minimizes collisions and energy consumption, it is not suitable for real-time applications as it decreases throughput. In addition, the cluster-based approach is not suitable for deep-water and ultra-deep water applications, as it requires a long time to transmit packets to the sink [13]. In contrast, the ad-hoc based approach allows nodes to transmit data whenever there is data to send. This approach suffers from high collisions and energy consumption, but improves throughput. Therefore, the ad-hoc based approach is more commonly used for real-time applications. The goal of the network design is to achieve low energy consumption while considering throughput, delay, and Packet Delivery Ratio (PDR). These performance factors are the most common metrics used to evaluate MAC protocols for UASNs. Avoiding collisions, adaptivity, fairness, scalability, and channel utilization are also important factors that influence the MAC protocols in UASNs [15]. The PS has a direct impact on network performance, such as throughput, delay, and energy consumption. The ideal PS depends on the NL (packets/sec), DR, and BER. Different underwater applications may require different network topologies. In this paper, we focus on ad-hoc multi-hop networks, as this is the most efficient approach to monitor the oil/gas pipeline.

12.2.1 Challenges of UASN Communication

The characteristics of the acoustic channel used by UASNs are extremely limited when compared to TWSNs due to the limited bandwidth, power, memory, long propagation delay, high Bit Error Rate (BER), and unreliable communication. The available bandwidth in UASNs can be up to 100 KHz, whereas TWSNs can reach up to 928 MHz. The propagation speed of UASNs is 1,500 m/s, while TWSNs utilize a propagation speed of 300,000,000 m/s. Sensor nodes in both UASNs and TWSNs are powered by limited batteries. However, transmitting packets between intended parties consumes higher energy in UASNs compared to TWSNs. In addition, sensors nodes in UASNs cannot be recharged or use solar energy while sensors nodes in TWSNs can be recharged and utilize solar energy [6]. Therefore, the network lifetime of sensor nodes in UASNs is shorter than those in TWSNs. The link between intended parties is more unstable due to the high BER and attenuation compared to TWSNs. Commonly, sensors in UASNs communicate over a long distance while TWSNs communicate over a short distance. Transmitting packets a longer distance requires higher power, which consumes higher energy. Thus, the network lifetime in UASNs is shorter compared to TWSNs. In Table 12.1, we have highlighted the major differences between UASNs and TWSNs.

Communication in UASNs is also affected by non-environmental factors, including limited bandwidth, multipath propagation, path loss, long variable delay, noise, and doppler spread [4, 6, 15]. These factors are discussed as follows:

- 1) **Limited bandwidth:** As demonstrated in Table 12.1, the available bandwidth is limited in UASNs compared to TWSNs. However, the available bandwidth in UASNs

Table 12.1 Characteristics of TWSNs vs. UASNs [16, 17]

Parameters	TWSNs	UASNs
Common communication modality	Radio waves	Acoustic waves
Propagation speed	300,000,000 m/s	1,500 m/s
Transmission range	10 m–100 m	Up to 10 Km
Frequency	908–928 MHz	10 Hz–100 KHz
Mobility of nodes	Application-based	Generally mobile
Reliability of links	Application-based	Low
Stability of links	Stable	Unstable
Localization	GPS supportive	GPS non-supportive
Node density	Dense	Generally sparse
Energy consumption	Low	High
BER	Moderate	High
Path loss	Low	High
Noise	Less impact	High impact

depends on communication ranges [18]. The higher the communication range, the lower the available bandwidth and vice versa [6]. The limited bandwidth can result in network congestion, which can result in higher end-to-end delay, packet loss ratio, and energy consumption due to the low transmission rate. Therefore, the limited bandwidth is considered a major issue in UASNs.

- 2) **Multipath propagation:** Multipath effects can occur due to variation of propagation speed, which results in a variance of propagation delay and wave reflections at the service level or ocean bottom.
- 3) **Path loss:** Path loss can occur due to geometric spreading and high attenuation. Geometric spreading occurs due to signal propagation. Attenuation is affected by distance and frequency. The higher the distance and frequency, the greater the attenuation. Hence, path loss can be decreased by minimizing the distance and utilizing high power for packet transmission [6].
- 4) **Long variable delay:** The propagation speed in UASNs is varied, as it depends on the underwater temperature, pressure, and salinity. Increasing any of the above factors results in higher propagation speed, which causes variable propagation delay.
- 5) **Noise:** Communication in UASNs is affected by human-made and ambient noises. These noises can cause interference to ongoing communication. As a consequence, packets must be retransmitted, thereby reducing the network lifetime.
- 6) **Doppler-spread:** Doppler-spread can occur due to communication range and mobile nodes. Due to the severe environment in UASNs, a sensor might move to a different position, which can cause doppler spread.

12.2.2 Network Architecture

Monitoring an underwater environment can be achieved in one-dimensional(1-D), two-dimensional (2-D), three-dimensional (3-D), or four-dimensional (4-D) architecture [19]. In 1-D architecture, each underwater sensor is a standalone network where each sensor collects and transmits data to the offshore infrastructure. The 2-D architecture is a set of underwater sensors, which can form a cluster and be placed at the same depth. The 3-D architecture is a set of underwater sensors in 2-D architecture, but underwater sensors are placed at different depths. The 4-D architecture is 3-D architecture plus Autonomous Underwater Vehicles (AUVs) with higher resources than underwater sensors.

12.2.3 Media Access Control Protocols

In our previous work [11], we review the MAC protocol approaches. In this paper, we focus on MAC protocols that are based on the hardware-based approach. The hardware-based approach further classifies the MAC protocol approaches into contention-free, contention-based, hybrid, and cross-layer. The contention-free MAC protocols are schedule-based where a sensor reserves the channel based on a unique code, frequency, or time. Time Division Multiple Access (TDMA) MAC (TMAC) protocol is a contention-free MAC protocol that allows sensor nodes to transmit packets at the beginning of the schedule-based time slot. The contention-based MAC protocols are clustered into the random access and handshaking MAC protocols. The random access approach further breaks down into without Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) and with CSMA/CA. The protocols that do not utilize the CSMA/CA allow sensor nodes to randomly transmit packets whenever they have data to send without any restrictions, such as Aloha [5]. In contrast, the protocols that utilize CSMA/CA allow sensor nodes to transmit packets only if the channel is idle within their transmission ranges, such as BroadcastMAC. In case the channel is busy within their transmission ranges, the sensor nodes back-off and sense the channel again at another time. The handshaking approach requires sensor nodes to transmit control packets prior to data transmissions to avoid collisions. The hybrid MAC protocols rely on both contention-free and contention-based approaches, such as Reservation-based MAC (RMAC). They take advantage of contention-free as it is more efficient to avoid collision while the contention-based is more appropriate to time-sensitive applications. The cross-layer MAC protocols share information about multiple layers to utilize network resources more efficiently and achieve high QoS. In this study, we chose to evaluate four different hardware-based MAC protocol approaches to determine which one is the most efficient for monitoring the oil/gas pipeline application. We also need to analyze the impact of using different underwater commercial modems on the common network performance required by such a time-sensitive application.

12.3 Related Work

One of the primary goals of designing an efficient and reliable MAC protocol for UASNs is to obtain the right balance between the Quality of Service (QoS) and an application's requirements. To meet an application's requirements in terms of network performance, developers must consider the envi-

ronmental factors, including water temperature, transmission range, salinity, and node mobility [20].

Besides environmental factors, developers must also consider the following ten non-environmental factors while designing an efficient MAC protocol for UASNs [15]: communication modality, network architecture, routing protocol, DR, hop length, NS, PS, NL, limited power, and antenna type.

Each one of these factors plays a vital role in enhancing the network performance for UASNs. However, some factors have a stronger impact than others. The network architecture has an impact on energy efficiency, throughput, delay, and PDR. A higher bandwidth in a routing protocol can result in a high end-to-end delay. The DR, NS, and protocol type are significant factors for throughput and energy consumption. Roy and Sarma [20] reported that the parameters that most significantly affect the performance of MAC protocols in UASNs are DR, hop length, and network topology, respectively. As DR increases, the energy consumption and throughput increase, but throughput drops when it surpasses a threshold value. Increasing the hop length increases delay and PDR along the network diameter, while throughput decreases. Increasing the NS increases the delay while decreasing energy consumption, throughput, and PDR. Increasing the NS while maintaining the same packet generation rate increases the throughput while decreasing the overhead and energy consumption. The PS has a significant impact on energy consumption, delay, and throughput. To determine the ideal PS, developers must consider BER, protocol attributes, DR, and NL. Underwater nodes have limited power, which has a direct impact on throughput, delay, and PDR. Different antennas have different designs, and the antenna size influences several network performance metrics. Table 12.2 summarizes the tradeoffs of the design considerations over two-hop in terms of the network performance.

Several mechanisms can be used to allow underwater nodes to communicate with one another. However, within each mechanism, many factors affect network performance.

Table 12.2 Tradeoffs of design considerations to underwater network performance over two-hop

Design factors/metrics	EC	Thpt	D	PDR	O
Communication modality	SI	SI	SI	SI	UI
Network architecture	SI	SI	SI	SI	UI
Higher bandwidth routing protocol	DP	UI	UI	UI	UI
Increasing data rate	DP	DP	IP	DP	UI
Increasing hop length	IP	IP	DP	IP	UI
Increasing network size	IP	IP	DP	IP	IP
Increasing packet size	SI	SI	SI	UI	UI
Limited power	UI	DP	DP	DP	UI
Antenna type	SI	SI	SI	SI	UI

EC: Energy Consumption, *Thpt*: Throughput, *D*: Delay, *O*: Overhead, *DP*: Directly Proportional, *IP*: Inversely Proportional, *SI*: Somewhat Impactful, *UI*: Undefined Impactful

Nasri et al. [14] have investigated the network topology considerations of legacy MAC protocols based on communication range and traffic load. The network topology, communication range, and traffic load are important factors to improve the network performance, but other factors must be considered, such as the location of underwater nodes. Therefore, Climent et al. [21] analyzed the research progress for underwater MAC protocols and required that the developers consider the location of sensor nodes and time synchronization in performance evaluation in order to have a more accurate reading of their performance. Time-sensitive underwater applications have specific network performance metrics, where non-time sensitive applications trade latency for network lifetime. Thus, Zenia et al. [22] have investigated several MAC protocols based on energy consumption and reliable communication. Also, authors in [21, 22] showed that these two factors play a critical role in resolving several issues of the MAC protocol in UASNs.

12.4 Factors Impact on Performance Metrics

In the previous section, several papers have highlighted that non-environmental factors, including DR, NS, PS, and NL play a vital role in resolving the issues of MAC protocols in UASNs and thus enhancing network performance [20–22]. Therefore, in this paper we comprehensively analyze the impact of different DRs, NSs, PSs, and NLs compared to the end-to-end delay, energy consumption, PDR, and collision rate of different MAC protocols. By knowing the relationship between the above non-environmental factors to the network performance, we can optimize the UASNs communication to meet the QoS of the underwater oil/gas pipeline application.

12.4.1 Underwater Commercial Modems

Underwater sensors have limited characteristics. These characteristics depend on the Underwater Acoustic Modems (UAMs) and therefore the network performance relies on these characteristics [23]. Each underwater commercial modem can be used for different underwater applications as some have higher DRs, but at the cost of increased energy consumption. In contrast, some underwater commercial modems have lower DRs, but at the cost of increased end-to-end delay. Therefore, selecting the appropriate underwater commercial modem is important to meet the QoS of the underwater application. Zia et al. [24] shows a comprehensive survey of all underwater acoustic modems available, including both commercial and research modems. Since we focus on a time-sensitive application, we need to

Table 12.3 Comparison the characteristics of commercial underwater acoustic modems [24]

Manufacturer	Model number	Range (m)	Data rate (bps)	Max. depth (m)	Operating frequency (kHz)	Transmission power (W)	Receiving power (W)	Standby power (W)
LinkQuest [25]	UWM4000	4,000	8,500	7,000	12.75–21.25	7	0.8	0.008
LinkQuest [25]	UWM1000	350	17,800	200	26.77–44.62	2	0.75	0.008

compare between two modems with different DRs, while minimizing energy consumption. Thus, we have selected the most efficient underwater commercial modems with different DRs. Table 12.3 shows the characteristics of the underwater acoustic modems that have been used in this study.

12.4.2 Data Rate

Different underwater applications have different monitoring requirements. Some applications require periodic monitoring at specific intervals, while others monitor only in case of irregular events. In addition, some applications require minimum delay, while others sacrifice delay for low energy consumption. In both cases, DR is an important factor to meet these requirements. Commonly, low DR results in high end-to-end delay, high collision rate, low PDR, and low energy consumption. In contrast, high DR reduces end-to-end delay and collision rate, which helps to achieve high PDR while the amount of energy consumption relies on the type of MAC protocol approach. Therefore, in Sect. 12.5 we simulate different DRs to analyze the impact of DR to the end-to-end delay, energy consumption, PDR, and collision rate.

12.4.3 Network Size

The NS has a strong impact on the network performance as the higher the number of nodes, the higher end-to-end delay while reducing the energy consumption [5]. Moreover, increasing the NS may increase or decrease the PDR depending on the MAC protocol approach, hop length, and network topology. In a multi-hop network the number of collisions is higher compared to a single-hop network and may reduce the PDR. Similarly, the chance of collisions is higher in an ad-hoc network compared to a clustered network.

12.4.4 Packet Size

Yildiz et al. [26] and Awan et al. [2] have shown that PS can greatly impact network performance. Selecting the proper PS can affect the results of end-to-end delay and energy consumption. However, the optimal PS relies on the MAC protocol characteristics, BER, NL, and DR [27]. Thus, in Sect. 12.5 we simulate different PSs to analyze the impact of

PS on the end-to-end delay, energy consumption, PDR, and collision rate.

12.4.5 Network Load

The NL represents the number of packets that can be sent over a specific interval. A high NL can increase the number of collisions as the underwater channel will experience high traffic. Hence, high NL can minimize the network lifetime and increase the end-to-end delay [27]. However, the impact of the NL depends on the MAC protocol characteristics, DR, NS, and PS.

12.5 Performance Evaluation

In this section, we simulate a 3D static network of a shallow underwater oil/gas pipeline over multi-hop by using several nodes at different depths and a single sink node at the surface level. We then analyze the behavior of underwater sensors that communicate in an ad-hoc based approach where each sensor transmits packets hop by hop in the network. Monitoring the underwater oil/gas pipeline requires minimum delay while maximizing PDR to avoid disaster [28]. In order to meet the QoS of this type of application, we investigate four different MAC protocol approaches discussed in Sect. 12.2 to determine which is the most efficient. To optimize the UASNs' communication, we also analyze the relationship between the most significant non-environmental factors, including DR, NS, PS, and NL and the network performance.

12.5.1 Simulation Setup

For this experiment, the VirtualBox software was used to set up a virtual machine running NS-2 on top of a Mint server. We have evaluated the MAC protocols discussed in Sect. 12.2 using an NS-2 based simulator for UASNs called Aqua-Sim. Based on our extensive investigation in [11], we evaluate the performance of these MAC protocols and utilize a simple network topology as shown in Fig. 12.1. In this network topology, we rely on the multi-hop structure to transmit packets between intended parties. In our simulation, the underwater sensors are static and randomly distributed. In Table 12.4, we illustrate the simulation parameters used in this study.

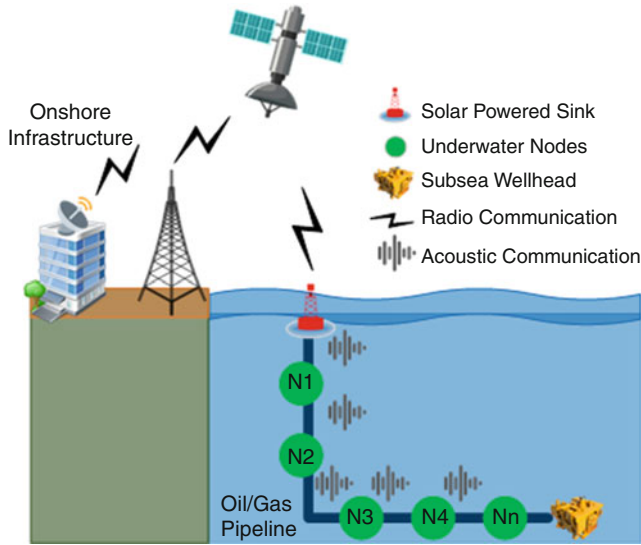


Fig. 12.1 Simulation network topology

Table 12.4 Simulation parameters

Parameters	Value
Radio propagation model	Underwater propagation
Channel	UnderwaterChannel
Routing protocol	Vectorbasedforward
Number of nodes	5 and 10
Simulation area	1000 m × 125 m
Transmission range	100 m
Simulation time	1500 s
Initial energy	10,000 J
Transmission power	Depends on UAM
Receiving power	Depends on UAM
Idle power	Depends on UAM
Data rate	8,500 bps and 17,800 bps
Network load	0.05, 0.1, 0.15, 0.2, 0.25 packets/sec
Data packet size	50, 100, 150, 200, 250 Bytes
Control packet size	5 Bytes
Type of traffic	Constant Bit Rate (CBR)

12.5.2 Performance Metrics

We measure the impact of different DRs, NSs, PSs, and NLs in terms of end-to-end delay, energy consumption, PDR, and total number of collisions. These performance metrics are described as follows:

- 1) End-to-End Delay (E2ED) is the total time each packet takes from source to destination plus processing delay. This is measured in seconds (sec). E2ED can be calculated using:

$$E2ED [sec] = \sum_{i=1}^{TS} (RT_i - ST_i) + P_d \quad (12.1)$$

Where, TS = total number of sent packets (TS); RT_i = received time of i th packet; ST_i = sending time of i th packet; P_d = processing delay.

- 2) Energy Consumption (EC) is the sum of energy consumed by all nodes, including transmitted power (Tx_p), receiving power (Rx_p), and idle power (Idl_p) to transmit packets between parties. This is measured in Joules (J). EC can be calculated using:

$$EC [J] = \sum_{i=1}^n (Tx_p + Rx_p + Idl_p) \quad (12.2)$$

Where, n = total number of nodes.

- 3) Packet Delivery Ratio (PDR) can be defined as the ratio of total number of received packets (TR) to the total number of sent packets (TS). PDR can be calculated using:

$$PDR = \left(\frac{TR}{TS} \right) \times 100 \quad (12.3)$$

- 4) Total Collision (TC) can be defined as the total number of sent packets (TS) that collide throughout the network. TC can be calculated using:

$$TC = \sum_{i=1}^{TS} SP_i \quad (12.4)$$

Where, SP_i = sent packet of i th.

12.5.3 Simulation Results

In this section, we present the network performance of different MAC protocols in UASNs using different DRs, NSs, PSs, and NLs. Due to the limited space, we only show the results of selected NLs of UWM1000 compared to UWM4000. We have selected these UAMs as the UWM1000 modem offers higher DR and lower EC compared to the UWM4000 modem. This helps to illustrate the impact of a higher DR on the network performance. The results in this section analyze the relationship between non-environmental factors to the network performance and compare the results of different UAMs. The non-environmental factors have different impact levels on network performance. Moreover, the impact level is different from one MAC protocol to another as each one has its own technique to transmit packets between intended parties. However, in all MAC protocols, the DR has the strongest impact, as it effects several performance metrics. Table 12.5 summarizes the relationship between the impact level of non-environmental factors and the network performance of UWM1000 compared to UWM4000. In this table, we show what affect a change on non-environmental variable has on the specific network performance metrics.

Table 12.5 Impact of non-environmental factors to underwater network performance over multi-hop of UWM1000 vs. UWM4000

Non-environmental factors/metrics	E2ED	EC	PDR	TC
Aloha				
Increasing data rate	IP	IP	DONF	DONF
Increasing network size	DP	DP	DONF	DP
Increasing packet size	DP	DP	DONF	DONF
Increasing network load	DP	DP	C	DP
BroadcastMAC				
Increasing data rate	IP	IP	DONF	DONF
Increasing network size	DP	DP	DONF	DP
Increasing packet size	DP	DP	DONF	DONF
Increasing network load	DP	DP	DONF	DP
RMAC				
Increasing data rate	IP	IP	DONF	N
Increasing network size	DONF	DONF	IP	N
Increasing packet size	DP	DP	DONF	N
Increasing network load	DONF	DONF	IP	N
TMAC				
Increasing network load	IP	IP	DONF	DONF
Increasing network size	DONF	DP	IP	DONF
Increasing packet size	DONF	DONF	DONF	DONF
Increasing network load	DONF	DP	DONF	DP

DP: Directly Proportional, IP: Inversely Proportional, C: Constant, DONF: Depend on Other Non-environmental Factors, N: None

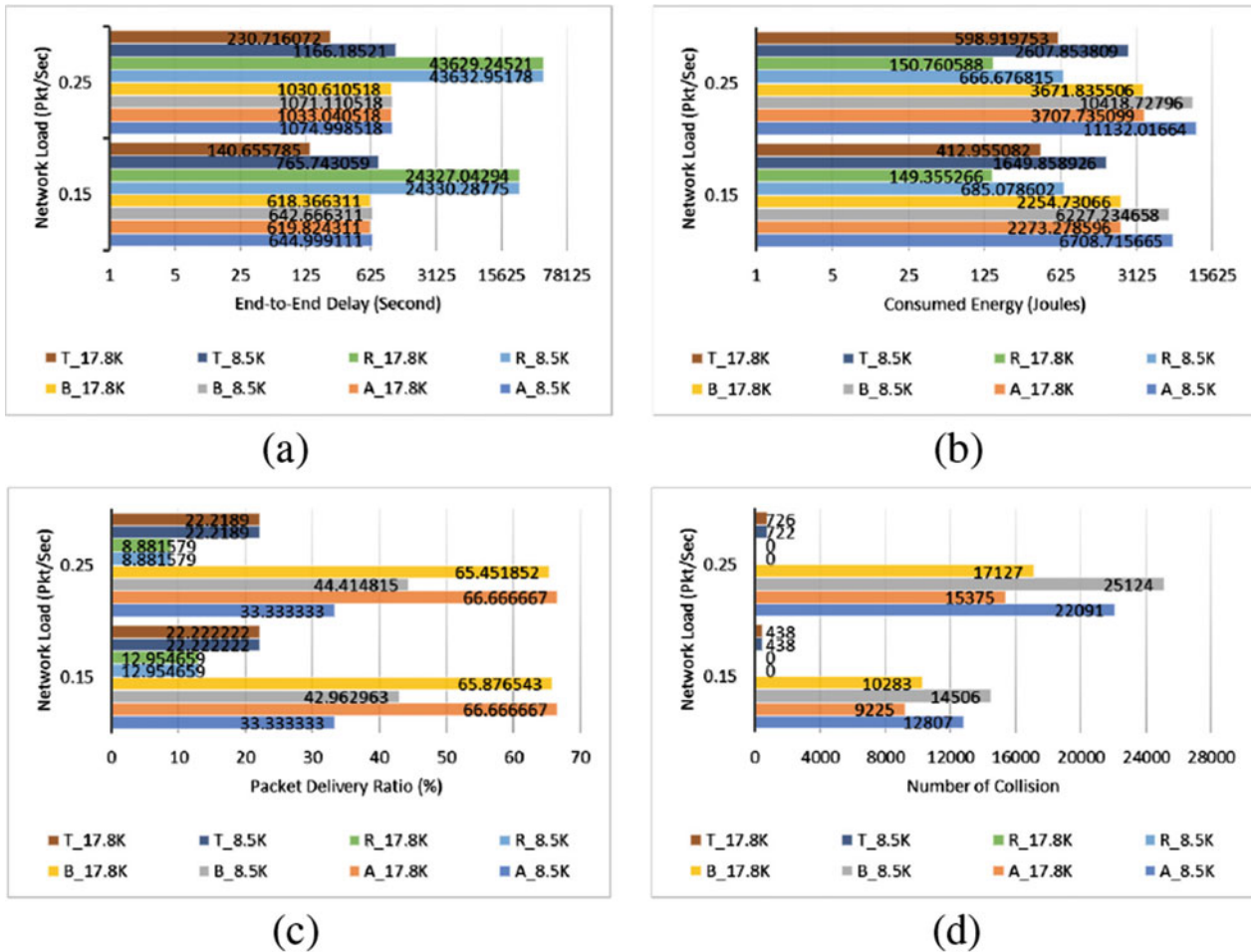
The tradeoff of some simulation results DONF as our experiment evaluates different non-environmental factors within each performance metric. For example, in Aloha the effect on PDR by increasing the DR depends on the NL. Higher NL decreases the PDR as more packets collided throughout the network. Hence, some non-environmental factors have multiple relationships with other factors. To analyze the results more accurately and show the impact compared to the performance metrics, we must use constant values of non-environmental factors. Figures 12.2, 12.3, 12.4, 12.5 compare each of the non-environmental factors discussed in Sect. 12.4 along with each performance metric discussed earlier.

In Fig. 12.2, we have evaluated different DRs using a large NS and 250 bytes as the PS. Clearly, increasing the DR should decrease the E2ED. However, the impact level of using different DRs in terms of E2ED is different from one MAC protocol to another. In Fig. 12.2a, the simulation results show that TMAC protocol is most strongly affected followed by Aloha, BroadcastMAC, and RMAC with about 82%, 4%, 4%, and less than 1%, respectively. Figure 12.2b shows the amount of EC where RMAC is the most affected followed by TMAC, Aloha, and BroadcastMAC with about 78%, 77%, 67%, and 65%, respectively. The use of RMAC results in the minimum EC, as there is no collision throughout the network. Although RMAC

and TMAC protocols consume less energy compared to Aloha and BroadcastMAC protocol, Fig. 12.2c shows that both Aloha and BroadcastMAC protocols have higher PDR compared to RMAC and TMAC protocols. Within an increased of DR, the PDR of RMAC and TMAC protocols maintain the same PDR due to packet transmission over multi-hop, while Aloha is the most affected followed by BroadcastMAC with 100% and 53%, respectively. Similarly, in Fig. 12.2d, increasing DRs does not impact RMAC and TMAC protocols, while BroadcastMAC decreases by 32% and Aloha by 30%. These results show that higher DRs have higher impact on the PDR followed by E2ED, EC, and TC, respectively.

Since using a higher DR results in a better performance, we utilize the DR of UWM1000 to analyze the results of different NSs, PSs, and NLs in Figs. 12.3, 12.4, 12.5. The simulation results in Fig. 12.3 are based on a PS of 250 bytes. Figure 12.3 shows the relationship between increasing NS and the network performance. In Fig. 12.3a, the simulation results reveal that by increasing the DR BroadcastMAC is most strongly affected followed by Aloha, RMAC, and TMAC with about 432%, 431%, 17%, and 1%, respectively. Similarly, Fig. 12.3b shows that the NS has a higher impact on Aloha followed by BroadcastMAC, RMAC and TMAC with about 261%, 247%, 41%, and 25%, respectively. In Fig. 12.3c, the PDRs of Aloha and BroadcastMAC protocols are higher than RMAC and TMAC protocols. Increasing the NS has higher impact on RMAC followed by TMAC, BroadcastMAC, and Aloha with about 57%, 56%, 13%, and 11%, respectively. Figure 12.3d shows that increasing NS has no effect on RMAC and TMAC protocols, while BroadcastMAC increases by up to 662% and Aloha by up to 584%. These results show that larger NS have higher impact on TC followed by E2ED, EC, and PDR, respectively.

We show in Fig. 12.4 the relationship between the PS and network performance. The simulation results in Fig. 12.4 are based on a NS of 10. Increasing the PS increases the E2ED of all MAC protocols. However, in Fig. 12.4a we show that TMAC protocol is most strongly affected followed by BroadcastMAC, Aloha, and RMAC with about 34,004%, 3%, 3%, and less than 1%, respectively. Figure 12.4b shows that increasing PS has the strongest effect on BroadcastMAC followed by RMAC, Aloha, and TMAC by 52%, 50%, 46%, and 37%, respectively. In Fig. 12.4c, the PDRs of Aloha and BroadcastMAC protocols are higher than RMAC and TMAC protocols. In addition, increasing the PS has no effect on RMAC and TMAC protocols. In contrast, increasing the PS has a higher impact on BroadcastMAC followed by Aloha with about 35% and 33%, respectively. Similarly, in Fig. 12.4d we show that increasing the PS has no effect on RMAC and TMAC protocols, while BroadcastMAC in-



A_8.5K: Aloha_DR_8.5K, B_17.8: BroadcastMAC_DR_17.8K, R_8.5K: RMAC_DR_8.5K, T_17.8K: TMAC_DR_17.8K

Fig. 12.2 Comparison of different data rates. (a) DR vs. E2ED (b) DR vs. EC (c) DR vs. PDR (d) DR vs. TC

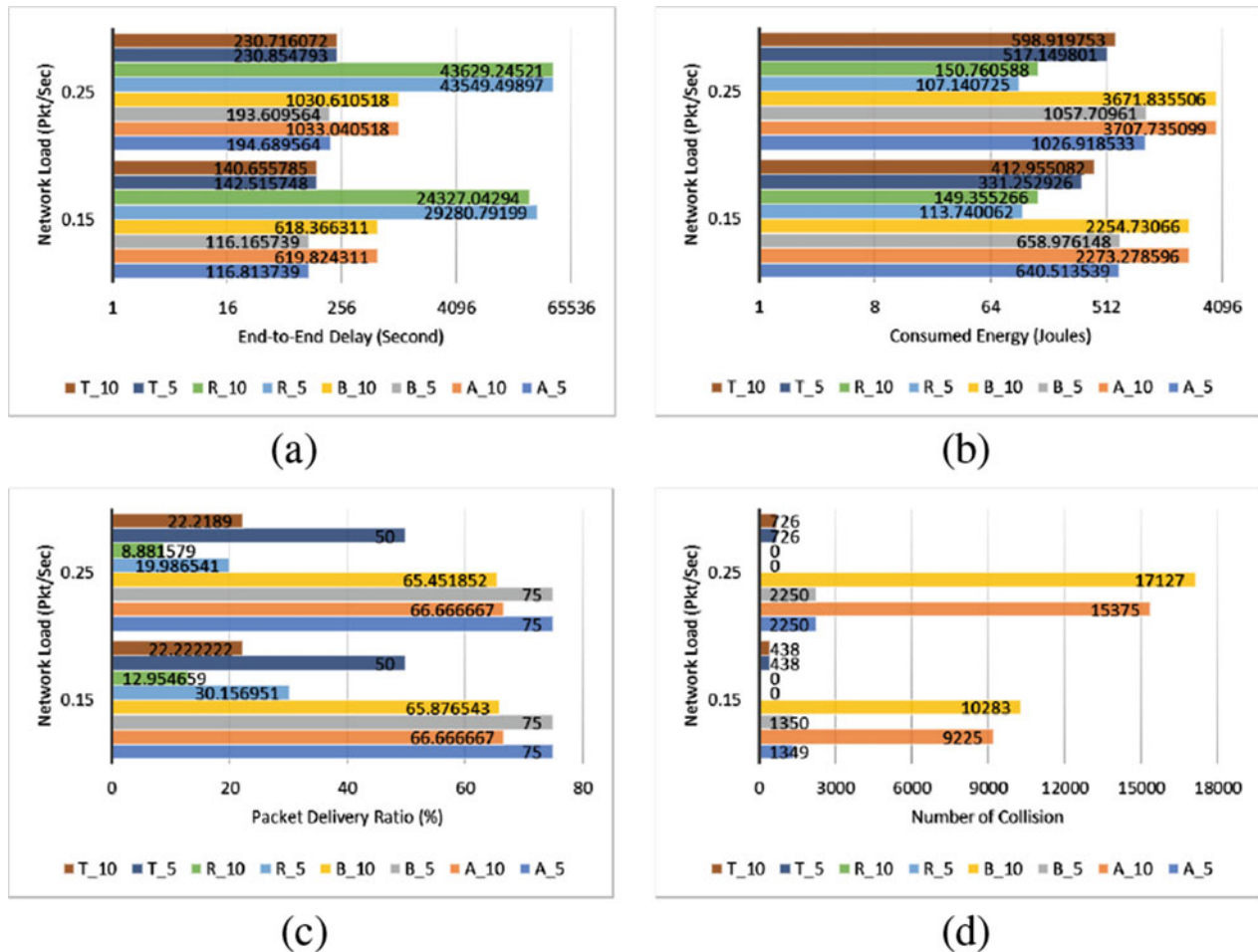
increases by up to 129% and Aloha by up to 105%. This illustrates that increasing PS has a higher impact on TC followed by EC, PDR, and E2ED.

In Fig. 12.5 we show the relationship between increasing the NL and the network performance metrics. The simulation results in Fig. 12.5 are based on a NS of 10 and PS of 250 bytes. In Fig. 12.5a, we show that RMAC protocol is most strongly affected followed by BroadcastMAC, Aloha, and TMAC with about 79%, 67%, 67%, and 64%, respectively. Figure 12.5b shows that the NL has the effect of increasing the performance of Aloha and BroadcastMAC by approximately 63%. TMAC performance was effected by 45% and RMAC by less than 1%. In Fig. 12.5c, the PDR remains the same on Aloha, while it decreases on all other MAC protocols. However, the RMAC protocol is the most strongly affected followed by BroadcastMAC, and TMAC

with about 31%, 1%, and less than 1%, respectively. We show in Fig. 12.5d that increasing NS is most strongly affected on Aloha and BroadcastMAC followed by TMAC with about 67% and 66%, respectively. These results show that increasing the NL have higher impact on E2ED, TC, EC, and PDR, respectively.

12.6 Conclusion

This paper highlights the background, issues, and challenges of underwater communication using UASNs. In order to efficiently monitor the oil/gas pipeline application, the developers must consider many factors that impact network performance. This study reveals that different MAC protocol approaches have different performance results. Therefore,



A_5: Aloha_NS_5, B_10: BroadcastMAC_NS_10, R_5: RMAC_NS_5, T_10: TMAC_NS_10

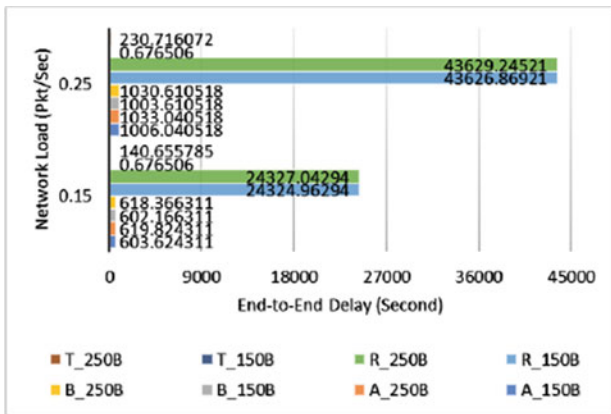
Fig. 12.3 Comparison of different network sizes. (a) NS vs. E2ED. (b) NS vs. EC. (c) NS vs. PDR. (d) NS vs. TC

each MAC protocol can be appropriate for certain underwater applications. Several papers have analyzed the performance of UASNs over single-hop and two-hop, whereas this paper provides a comprehensive analysis of non-environmental factors over multi-hop. We focus on the relationship between the non-environmental factors and the network performance to determine the most efficient MAC protocol approach for such a time-sensitive application. The DR has the strongest impact on the network performance. However, the impact level varies from one MAC protocol to another. Thus, analyzing the effect of non-environmental factors that affect the network performance helps developers to be aware of the tradeoffs and hence to optimize the UASNs communication to meet the QoS of an underwater application.

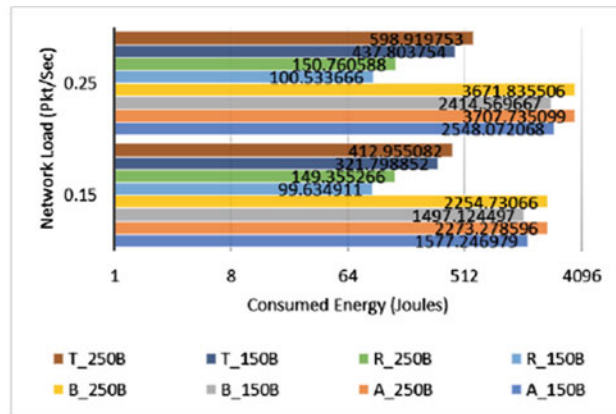
Our extensive simulation results have shown that the BroadcastMAC followed by Aloha, TMAC, and RMAC re-

spectively, are the most suitable for a time-sensitive application. Accordingly, we conclude that the contention-based random access approach is the most suitable for monitoring the oil/gas pipeline application. The contention-based MAC protocols show that the non-environmental factors that have the strongest impact on delay are NS followed by NL, DR, and PS, respectively. Further investigation of the UASNs using other commercial modems that offer higher DRs is needed to enhance the network performance of UASNs. In addition, a comprehensive analysis requires analyzing other performance metrics, including throughput, routing overhead, fairness-index, and jitter.

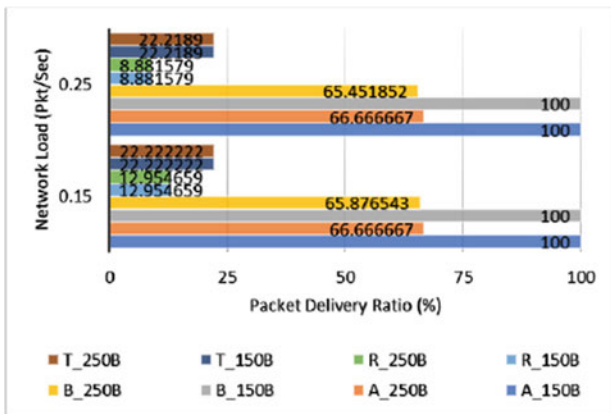
Acknowledgments This work was partially supported by the Department of Computer Science and Graduate School of University of Colorado Colorado Springs, King Faisal University, Saudi Arabian Cultural Mission (SACM) in USA, and University of Tabuk.



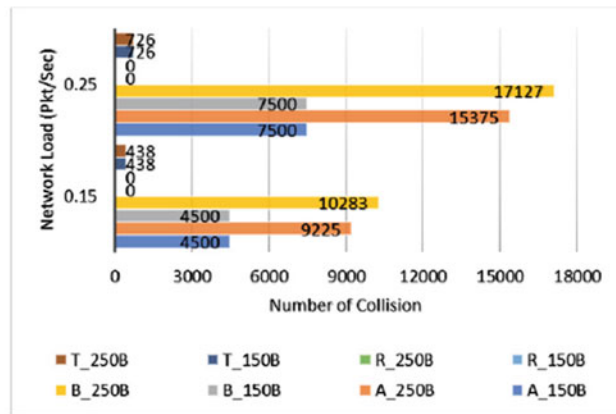
(a)



(b)



(c)



(d)

A_150B: Aloha_PS_150B, B_250B: BroadcastMAC_PS_250B, R_150B: RMAC_PS_150B, T_250B: TMAC_PS_250B

Fig. 12.4 Comparison of different packet sizes. (a) PS vs. E2ED. (b) PS vs. EC. (c) PS vs. PDR. (d) PS vs. TC

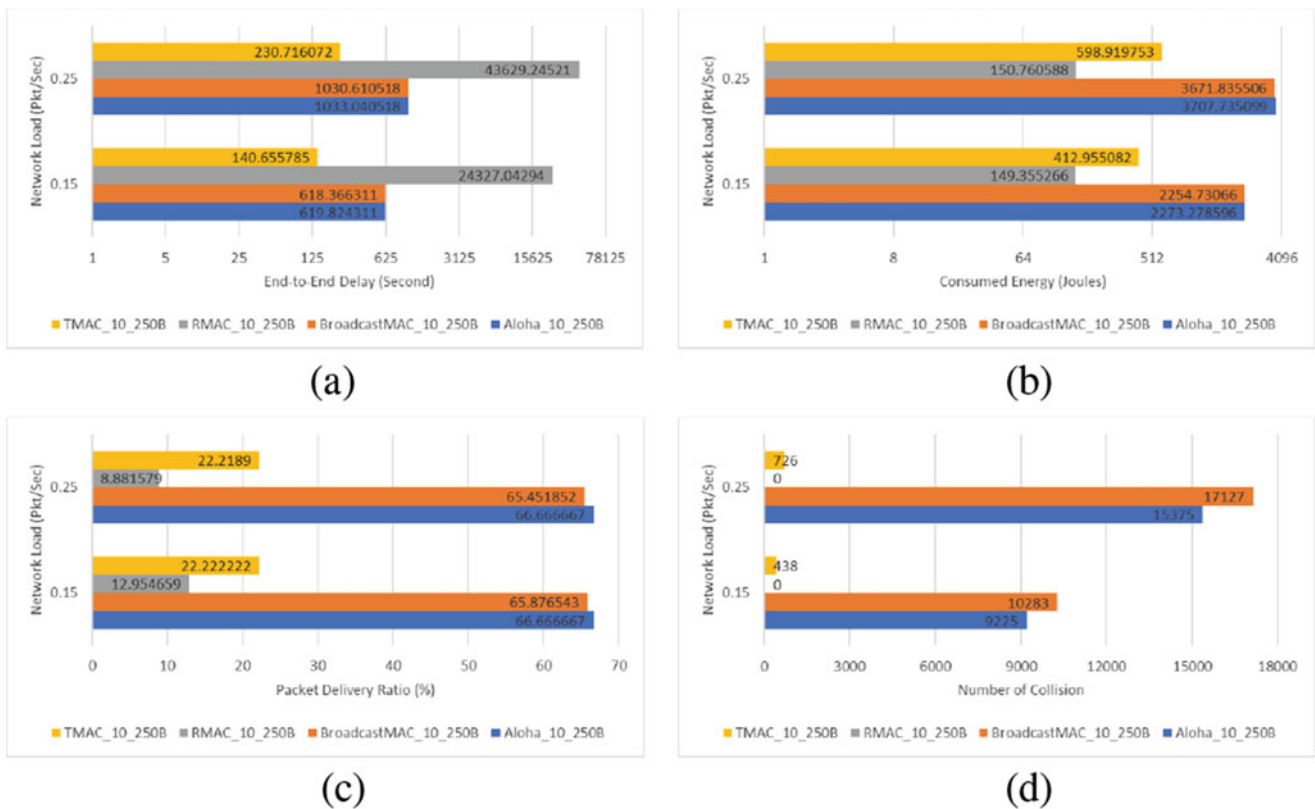


Fig. 12.5 Comparison of different network loads. (a) NL vs. E2ED. (b) NL vs. EC. (c) NL vs. PDR. (d) NL vs. TC

References

- L.K. Narayanan, S. Sankaranarayanan, Multi-agent based water distribution and underground pipe health monitoring system using iot, in *Information Technology-New Generations* (Springer, 2019), pp. 395–400
- K.M. Awan, P.A. Shah, K. Iqbal, S. Gillani, W. Ahmad, Y. Nam, Underwater wireless sensor networks: A review of recent issues and challenges. *Wireless Commun. Mobile Comput.* **2019** (2019)
- Offshore Oil Production in Deepwater and Ultra-deepwater is Increasing, U.S. Energy Information Administration - EIA - Independent Statistics and Analysis, 28-Oct-2016. [Online]. Available: <https://www.eia.gov/todayinenergy/detail.php?id=28552>. [Accessed: 05-Oct-2020]
- H. Khan, S.A. Hassan, H. Jung, On underwater wireless sensor networks routing protocols: A review. *IEEE Sensors J.* **1748**(c), 1–1 (2020)
- V. Casares-Giner, T.I. Navas, D.S. Flórez, T.R. Vargas, End to end delay analysis in a two tier cluster hierarchical wireless sensor networks, in *Information Technology-New Generations* (Springer, 2018), pp. 149–158
- I.F. Akyildiz, D. Pompili, T. Melodia, Challenges for efficient communication in underwater acoustic sensor networks. *ACM SIGBED Rev.* **1**(2), 3–8 (2004)
- P. Lacovara, High-Bandwidth Underwater Data Communication System (2020, Apr. 14). Patent US 10,623,110 B2 [Online]. Available: <https://patents.google.com/patent/US10623110B2>
- C. Youngbaull, D. Ganger, A. Mora, A. Richa, J. Zhang, C. Zhou, X. Hu, Underwater Multi-Hop Communications Network (2016, May 12). Patent US 2016/0134433 A1 [Online]. Available: <https://patents.google.com/patent/US20160134433A1>
- I.F. Akyildiz, P. Wang, Z. Sun, Realizing underwater communication through magnetic induction. *IEEE Commun. Mag.* **53**(11), 42–48 (2015)
- M. Rhodes, D. Wolfe, B. Hyland, Communications System, (2019, Aug. 15). Patent US 2019/0253156 A1 [Online]. Available: <https://patents.google.com/patent/US20190253156A1>
- A. Al Guqhaiman, O. Akanbi, A. Aljaedi, C.E. Chow, A survey on MAC protocol approaches for underwater wireless sensor networks. *IEEE Sensors J.* 1–16 (2020)
- S. Jiang, State-of-the-art medium access control (MAC) protocols for underwater acoustic networks: A survey based on a MAC reference model. *IEEE Commun. Surv. Tutorials* **20**(1), 96–131 (2018)
- M. Sharif-Yazd, M.R. Khosravi, M.K. Moghimi, A survey on underwater acoustic sensor networks: Perspectives on protocol design for signaling, MAC and routing. *J. Comput. Commun.* **05**(05), 12–23 (2017)
- N. Nasri, A. Kachouri, L. Andrieux, Surveys of design considerations for underwater networks. *Int. J. Inf. Commun. Technol.* no. February 2010, 1–5 (2009)
- A. Roy, N. Sarma, Effects of various factors on performance of MAC protocols for underwater wireless sensor networks. *Mater. Today Proc.* **5**(1), 2263–2274 (2018)
- S. Sahana, K. Singh, R. Kumar, S. Das, A review of underwater wireless sensor network routing protocols and challenges. *Next Gen. Networks* 505–512 (2018)
- D.N. Sandeep, V. Kumar, Review on clustering, coverage and connectivity in underwater wireless sensor networks: A communication techniques perspective. *IEEE Access* **5**, 11176–11199 (2017)
- F. Campagnaro, R. Francescon, P. Casari, R. Diamant, M. Zorzi, Multimodal underwater networks: Recent advances and a look

- ahead, in *ACM International Conference on Underwater Networks & Systems - WUWNet '17*, pp. 1–8, 2017
19. E. Felemban, F.K. Shaikh, U.M. Qureshi, A.A. Sheikh, S.B. Qaisar, Underwater sensor network applications: A comprehensive survey. *Int. J. Distrib. Sensor Networks* **2015**, (2015)
 20. A. Roy, N. Sarma, Factors affecting MAC protocol performance in underwater wireless sensor networks. *Int. J. Comput. Appl.* **169**(5), 36–41 (2017)
 21. S. Climent, A. Sanchez, J.V. Capella, N. Meratnia, J.J. Serrano, Underwater acoustic wireless sensor networks: Advances and future trends in physical, MAC and routing layers. *Sensors (Switzerland)* **14**(1), 795–833 (2014)
 22. N.Z. Zenia, M. Aseeri, M.R. Ahmed, Z.I. Chowdhury, M. Shamim Kaiser, Energy-efficiency and reliability in MAC and routing protocols for underwater wireless sensor network: A survey. *J. Network Comput. Appl.* **71**, 72–85 (2016)
 23. S. Sendra, J. Lloret, J.M. Jimenez, L. Parra, Underwater acoustic modems. *IEEE Sensors J.* **16**(11), 4063–4071 (2016)
 24. M.Y.I. Zia, J. Poncela, P. Otero, *State-of-the-Art Underwater Acoustic Communication Modems: Classifications, Analyses and Design Challenges*, vol. 0123456789 (Springer US, 2020)
 25. Link Quest Telecom Ltd., LinkQuest Underwater Acoustic Modems Features. [Online]. Available: <http://link-quest.com/html/models1.htm>. [Accessed: 11-Oct-2020]
 26. H.U. Yildiz, V.C. Gungor, B. Tavli, Packet size optimization for lifetime maximization in underwater acoustic sensor networks. *IEEE Trans. Ind. Inf.* **15**(2), 719–729 (2019)
 27. M. Yigit, H.U. Yildiz, S. Kurt, B. Tavli, V.C. Gungor, A survey on packet size optimization for terrestrial, underwater, underground, and body area sensor networks. *Int. J. Commun. Syst.* **31**(11), 1–28 (2018)
 28. W.Z. Khan, M.Y. Aalsalem, W. Gharibi, Q. Arshad, Oil and gas monitoring using wireless sensor networks: Requirements, issues and challenges, in *International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pp. 31–35, 2016

Part III

Cybersecurity II

Parallelized C++ Implementation of a Merkle Tree

13

Andrew Flangas, Autumn Cuellar, Michael Reyes, and Frederick C. Harris Jr.

Abstract

Merkle trees are primarily known for being an attribute found in blockchain technology. They are used for encrypting data by hashing values multiple times to avoid incidents such as hash collisions, or the successful guessing of hash values. Merkle trees are not only a useful feature found on the blockchain but in the field of Cyber Security in general. This paper outlines the process of implementing a Merkle tree as a data structure in C++ and then parallelizing it using OpenMP. The final result is a Merkle tree password storing program with reduced running-time and the ability to operate on multiple processors. The validity of this program is tested by creating a Merkle tree of the correct passwords, storing the value of the root node, and then building a second tree where a single incorrect password is stored within that tree. The two trees are passed through an audit function that compares the root nodes of the two trees. If they are different, then the tree in question has been tampered with.

Keywords

Merkle tree · Parallel programming · Serial programming · Hash · OpenMP · Processors · Running-time · Root node · Cyber security · MD5 · UPC · MPI · Password · Data structure · Login data · Threads · Blockchain · Performance

A. Flangas · A. Cuellar · M. Reyes · F. C. Harris Jr. (✉)
 Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA
 e-mail: andrewflangas@nevada.unr.edu; acuellar24@nevada.unr.edu; michaelreyes@nevada.unr.edu; fred.harris@cse.unr.edu

13.1 Introduction

In order to ensure security in a rapidly changing field of technology, one must use techniques that are on the frontier of said field. A Merkle tree is an attribute primarily associated with blockchain, but it can also be used in the field of Cyber Security, in general, in ways such as storing sensitive data. By implementing a Merkle tree in C++, we are hoping to disassociate it from blockchain and use it primarily as a tool for Cyber Security purposes. This tool will then be parallelized for optimization aspects such as reduced running time, as well as the ability to run on multiple processors. The Merkle tree will also be implemented as a data structure for a program that will be able to test the efficacy and correctness of this tool. The Merkle tree data structure will be implemented twice: a serial version of the program and then a parallelized version in order to analyze the differences in the running-time and the number of processors being used.

The method that is used to parallelize the Merkle tree data structure is the application programming interface OpenMP [1], due to its simple and flexible interface for creating parallel programs in multiple languages including C++. The type of application that is used to test the Merkle tree data structure is that of a system for storing login data, in which the Merkle tree is used to hash and store the user passwords. In order to properly test the validity of this program, the known passwords for the login system will be fed into the application which will be the nodes for the construction of the Merkle tree. A second tree will also be constructed in which a single password from the original tree is altered. The root nodes of the two trees will then be passed through an audit function, which will compare the two roots to see if they are the same. If one of the roots is different, then it proves the tree in question has been tampered with.

The running-times and number of processor threads being used for both the serial and OpenMP versions of the program will be output to the terminal as well.

The rest of this paper is structured as follows: Merkle Trees and Parallelism is described in Sect. 13.2. Login Application Implementation is presented in Sect. 13.3, Performance Results are discussed in Sect. 13.4, and Conclusions and Future Work are covered in Sect. 13.5.

13.2 Merkle Trees and Parallelism

13.2.1 Merkle Trees

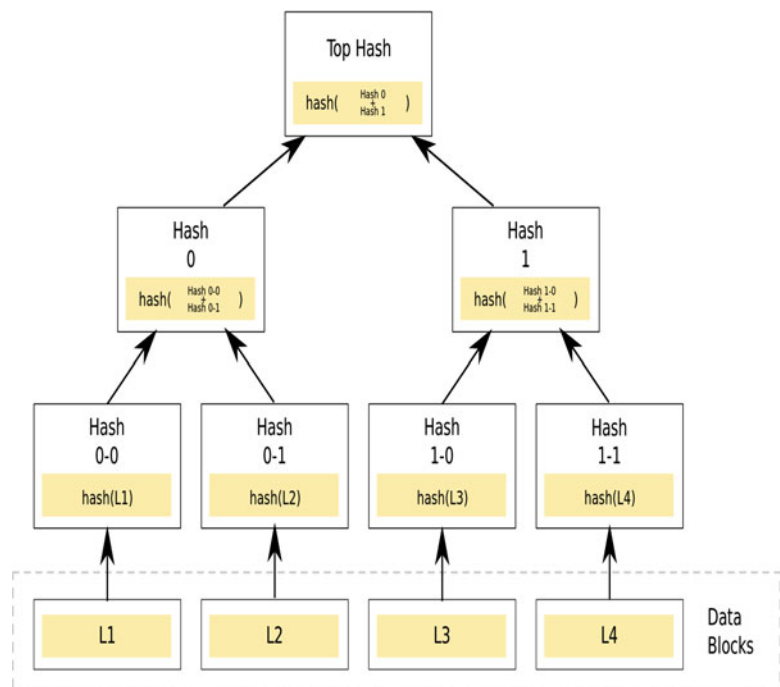
A Merkle tree is a data structure that combines binary trees and hash tables into one. Each node of the tree is a hash function of the combined children hash functions. The leaves are just a hash of the input value received while the root node is a hash of the whole tree, the structure can be seen in (Fig. 13.1). This tree can be very helpful in cyber-security applications. The root node's value is known based on the given input values. Any changes to the inputs, resulting from an attack, will cause the root node to change suddenly. This change will be detected which can allow users to determine the system as corrupt. Merkle trees are primarily used on blockchain applications, where the data is hashed using a Merkle tree and then the root node hash of the tree is added to a block. This allows for increased security measures when encrypting data that are not available when using traditional hashing methods.

Imagine a scenario of an online sealed-bid auction, where bidders simultaneously submit sealed bids to the auctioneer, so that no bidder knows the amount the other auction participants are bidding. Assuming the bid amounts of the participants are stored using traditional hashing methods, there is a way for one of the participants to easily figure out the other prices. All the malicious participant would have to do is guess the other bid prices and then run those values through a typical md5, sha1, or other traditional hashing algorithms to see if the hash values line up with the ones being stored. If the hashes do line up and the malicious participant was able to discover that the highest bid hash being stored is 30 dollars, then he or she would simply need to bid 31 dollars to win the auction. A Merkle tree prevents this scenario from happening by combining the hashes of the other bids into multiple levels of hashes until it reaches a root node, which adds extra encryption.

13.2.2 OpenMP

OpenMP is a C++ library that helps parallelize shared-memory programs. This library is comprised of library routines, environment variables, and compiler directives that influence the behavior of the running time. It does this through the use of threads and gives programmers a simple yet flexible interface for developing parallel programs. The main distinction between OpenMP and its close counterpart Message Passing Interface (MPI), is that OpenMP is used for parallelism within a multi-core node, whereas MPI is

Fig. 13.1 The structure of a Merkle tree [2]



used for parallelism between nodes. The one that will most likely see more favorable results in terms of decreasing the running-time is MPI over OpenMP, however, MPI is more complex and not as intuitive as OpenMP.

One of the main reasons OpenMP was used for this project over MPI was to prove that by parallelizing a Merkle tree using the most straight-forward method, will show a significant decrease in the overall running time of the program. OpenMP will be an efficient way to parallelize a Merkle tree program by splitting up the workload on a single machine. Each node can have its own thread to calculate the value at that specific node. This will be a quick computation to do because the thread can easily read the children node's values to use in the parent node's hash function. Locks will need to be used so valid data is used at every level of the tree. The parallelization of the Merkle tree will be done using the OpenMP pragma operations on the primary region of the code that heavily influences the running time of the program.

13.2.3 Unified Parallel C

Although Unified Parallel C or UPC++ was not used in this project, it has the potential to show promising results in regards to future work and is thus worth mentioning. UPC is a high-performance computing extension of the C programming language that can be used for large-scale parallel machines. UPC utilizes a shared global address space along with distributed memory that gives the programmer the ability to use a single shared, partitioned address space. This allows for variables to be directly read and written by any processor, but the variables are also physically correlated with a single processor.

The amount of parallelism is fixed at the startup time of the program due to the use of a single program, multiple data computation models, which usually results in a single thread of execution per processor. UPC combines the control over the data layout and performance of the message passing programming paradigm, with the programmability advantages of the shared memory programming paradigm which makes it a powerful tool for parallel programmers. This allows for the workload of a program to not only be split between other processors, but also different machines with their own sets of processors. Out of all the methods mentioned so far, UPC++ would probably give the best results in terms of running-time and processing power compared to the others.

13.2.4 Related Work

There is a method described in [3] that differentiates from the current standard of using naive locking for Merkle tree updates of the entire tree. This method is known as Angela

and is a distributed and concurrent sparse implementation of a Merkle tree. The method is distributed by utilizing Ray onto Amazon EC2 clusters, and then retrieves and stores the state using Amazon Aurora. The main aspect that Angela is motivated by is Google Key Transparency, which comes in direct inspiration from its underlying Merkle tree known as Trillian. Angela publishes a new root after some amount of time after assuming that a large number of its 2256 leaves are empty, which is the same task offered by Trillian. The approaches used by Google Trillian and the concurrent algorithm offered by Angela are compared, which shows nearly a 2x speedup.

Some other related research is stated in [4] which claims to be the first to provide complete, succinct, and recursive sparse Merkle tree definitions, along with related operations. These definitions show, when applied, that efficient space-time trade-offs for different caching strategies are enabled. It is also shown that utilizing SHA-512/256 to generate verifiable audit paths to prove (non-)membership, is done in nearly constant time which is less than 4ms. These results were concluded despite there being a limited amount of cache space, as well as there being a minimal effort of complete security embedded in the multi-instance setting. The size of the cache structure was smaller than the underlying data structure that was the target for authentication.

The paper [5] discusses hash functions derived from three modes of operation considering an inner Variable-Input-Length function. The inner function mentioned can be a sponge-based hash function, or a prefix-free MD and single-block-length(SBL) hash function. This paper discusses numerous techniques used for optimization purposes pertaining to developing parallel hash functions derived from trees in which all the leaves possess the same depth. The first result is comprised of a scheme that optimizes the topology of the tree to decrease the running time. The second result shows that the number of required processors can be minimized by slightly modifying the corresponding tree topology, without affecting the optimal running time. Therefore, this technique proves to reduce not only the running time but the number of required processors as well.

The hardware cost of implementing hash-tree based verification of untrustworthy external memory via a high-performance processor is reviewed in [6]. Certified program execution can be a result of this verification enabling these types of applications. Multiple schemes are displayed offering different integration levels that are between the on-processor L2 cache and the hash-tree machinery. A set of simulations also display the best version of the performance overhead that is less than 25 percent as a result of these methods. This is a significant decrease from the naive implementation, which normally presents a 10x overhead result.

Authenticated Data Structures (ADS) are discussed at length in [7], which defines an ADS to be data structures

whose operations can be carried out by an untrusted prover. This results in a verifier being able to efficiently check the authenticity of these operations. To create this scenario, the prover produces a compact proof which is then checked by the verifier, along with the results of each operation. Therefore, ADS supports the processing of tasks to untrusted servers without having to worry about the loss of integrity of the data, as well as outsourcing data maintenance. This paper also introduces a generic method that uses a simple extension to a programming language similar to what is used in machine learning algorithms, with which one can program authenticated operations over any data structure that is defined by standard-type constructors.

13.3 Password Storing Application Implementation

13.3.1 Program Structure

The first part of the implementation procedure was to build the serial version of the program. This entailed creating two data objects: a Merkle tree class and a Node struct. The class members of the Merkle tree include a Node pointer variable, as well as a printTree and deleteTree function. Included in the header file of the Merkle tree class is the declaration of the audit function that is used to detect whether the tree has been tampered with. The members of the Node struct include a string variable to hold the hash values, two Node pointers for the parent nodes, and lastly a function that passes a string to hold as data, that being the passwords. The main file is set up with the leaves of the tree declared as a vector of Node pointers and a Merkle tree pointer is used to build the tree of leaves. Using a for loop, the data is assigned and hashed to the parent nodes in the leaves vector. The output of a simplified built tree can be seen in (Fig. 13.2) along with the result of the audit function.

13.3.2 Program Functionality

The program operates by reading in strings of passwords from a file containing at max, 100,000 different passwords. It stores these passwords as the data that is used for the leaves of the Merkle tree. The passwords are hashed using the md5 hash library, after which the tree is fully constructed and output to the screen. This is the process for building one of the trees, but in order to test the validity of the program, two trees were constructed. The first tree is used to store all of the correct password information taken directly from the file whereas the second tree reads in the same passwords, except for altering one of the password values. The two root nodes

```

        e1
        5f
    04
        25
        d8
    91
9e
        25
        82
    d1
        81
        96
    f3
    04
65
The tree creation time is :2.2e-05
        e1
        5f
    04
        25
        d8
    91
9e
        25
        82
    d1
        81
        7a
    be
    dd
49
This tree may have been tampered with. Check data

```

Fig. 13.2 The output of the first two characters of the Md5 hashes for a version of the Merkle tree program with a small number of nodes. The output also displays the result of the audit function

of the trees are then compared using the audit function, which compares the second tree, that being the tree in question, with the first tree which is already known to be correct. If for any reason the root hashes differ, then it can be assumed that one of the password values is incorrect and the tree then becomes obsolete. A separate program was also included in a header file to keep track of the running times and print them to the screen.

13.3.3 Making the Program Parallel

The primary region of the program that influences the running time the most is located in the file where the Merkle tree class is defined. The differences between the parallel and serial implementations can be shown in Figs. 13.3 and 13.4. The main difference between the two implementations is the use of OpenMP in the parallel implementation, in which two pragma commands are utilized. The pragma omp parallel command is used to make the program parallel, and declare multiple threads to be run on a certain number of processors. The pragma omp for order operation is then used on the for

```

for (unsigned int l = 0, n = 0; l < blocks.size(); l = l + 2, n++) {
    if (l != blocks.size() - 1) { // checks for adjacent block
        nodes[n] = new Node(md5(blocks[l]->hash_val + blocks[l+1]->hash_val));
        nodes[n]->mom = blocks[l]; // assign children
        nodes[n]->dad = blocks[l+1];
    } else {
        nodes[n] = blocks[l];
    }
}

//std::cout << "\n";
blocks = nodes;
nodes.clear();
}

this->root = blocks[0];
}

```

Fig. 13.3 The serial implementation used in defining the Merkle tree

```

#pragma omp parallel
{
    #pragma omp for ordered
    for (unsigned int l = 0, n = 0; l < blocks.size(); l = l + 2, n++) {
        if (l != blocks.size() - 1) { // checks for adjacent block
            nodes[n] = new Node(md5(blocks[l]->hash_val + blocks[l+1]->hash_val));
            nodes[n]->mom = blocks[l]; // assign children
            nodes[n]->dad = blocks[l+1];
        } else {
            nodes[n] = blocks[l];
        }
    }
}

//std::cout << "\n";
blocks = nodes;
nodes.clear();
}

this->root = blocks[0];
}

```

Fig. 13.4 The parallel implementation used in defining the Merkle tree

loop to further optimize the parallelization of the program. The results of the two implementations can be seen in the next section.

13.4 Performance Results

To test the difference in running time between the parallel and serial implementation of the program, different test cases

```

[mreyes98@login006 final]$ ./main
The tree creation time is :0.00241

```

Fig. 13.5 The output of the running time for the serial program when fed 1,000 passwords

```

The tree creation time is :0.001867
The tree creation time is :0.001816
The tree creation time is :0.001811
The tree creation time is :0.001816
The tree creation time is :0.001893

```

Fig. 13.6 The output of the running time for the parallel program when fed 1,000 passwords

```

[mreyes98@login006 final]$ ./main
The tree creation time is :0.018375

```

Fig. 13.7 The output of the running time for the serial program when fed 10,000 passwords

```

The tree creation time is :0.020607
The tree creation time is :0.020549
The tree creation time is :0.017677
The tree creation time is :0.019503
The tree creation time is :0.020123

```

Fig. 13.8 The output of the running time for the parallel program when fed 10,000 passwords

```

[mreyes98@login006 final]$ ./main
The tree creation time is :0.192599

```

Fig. 13.9 The output of the running time for the serial program when fed 100,000 passwords

```

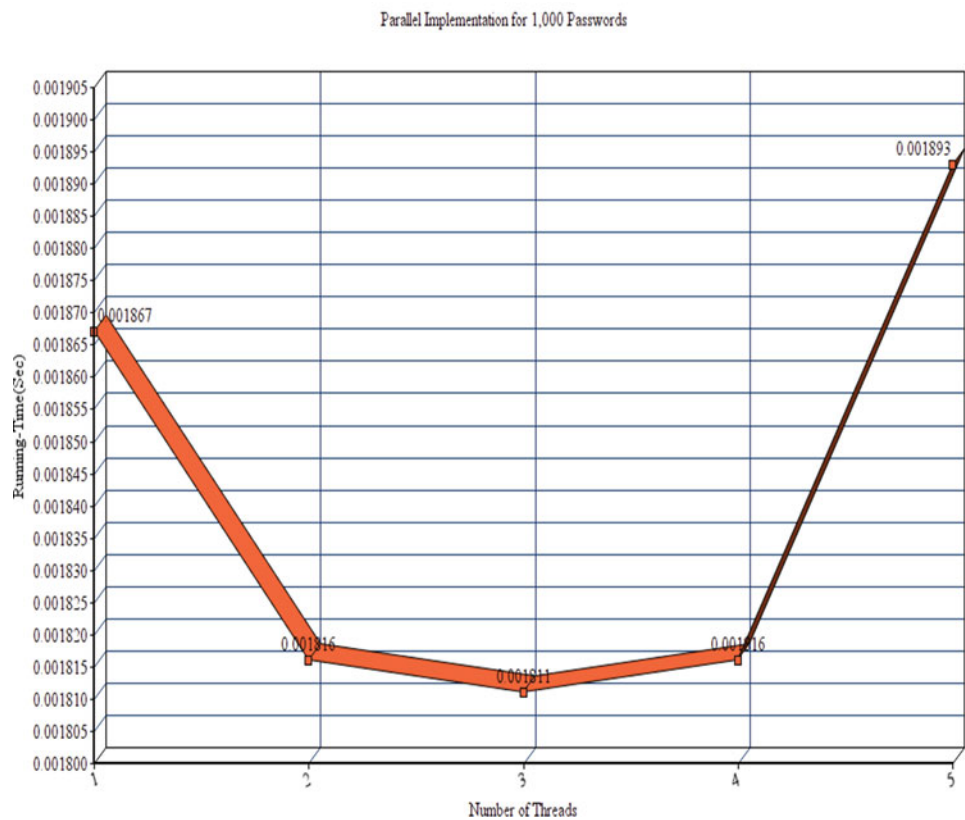
[mreyes98@login006 final]$ ./testFile
The tree creation time is :0.185375
The tree creation time is :0.176204
The tree creation time is :0.171894
The tree creation time is :0.176826
The tree creation time is :0.178371

```

Fig. 13.10 The output of the running time for the parallel program when fed 100,000 passwords

were executed on the Bridges supercomputer. There were three different test cases in which the number of passwords fed to the program was administered. The number of passwords fed to both the serial and parallel program for the first test case was 1,000 which can be shown in Figs. 13.5 and 13.6, 10,000 for the second which can be shown in Figs. 13.7 and 13.8, and 100,000 for the third which can be shown in Figs. 13.9 and 13.10. The reason for choosing this method to test the differences in running time was to illustrate the effectiveness of the parallelization which could be more easily observed when increasing the number of nodes in the tree.

Fig. 13.11 Line graph of the running-time for the parallel implementation using 1,000 passwords



The results for the 1,000 password test case show that there is a significant decrease in the running-time in the parallel implementation when run on all the processors by a factor of .0001. When tested using 10,000 passwords, however, there seemed to only be a decrease in the running-time in the parallel implementation when split between three threads, which was by a factor .001. The final case when tested using 100,000 passwords showed a significant decrease in running-time in the parallel implementation when running all processor threads by a factor of .01. These results show significant decreases in running-time when tested with the parallel implementation in all cases in at least one, or all of the processor thread options. All of the parallel implementation results show that the fastest running-time is achieved when the workload is split between three threads.

However, significant does not mean a large decrease in this context, it only means that it decreased enough to show that there were some optimization benefits to using OpenMP. A line graph of the running-times for each of the parallel implementations can be viewed in Figs. 13.11, 13.12, and 13.13. The Merkle tree data structure is a complex target to optimize due to the already quick processing capabilities inherent in its structure. In retrospect, perhaps a method involving the use of UPC++ would prove to be a better option for decreasing the running-time even more, due to its ability to split up the workload between different machines rather than several processors on the same machine.

13.5 Conclusions and Future Work

13.5.1 Conclusions

These implementations of a Merkle tree, both the parallel and serial versions, show that Merkle trees can be used with success for certain aspects unrelated to blockchain applications. First, the serial version of the Merkle tree was built and tested three times with passwords ascending in quantity with each iteration, and the results of the running-time were recorded. The method of testing involved two Merkle trees, one that held the correct list of passwords and another with the same list with one of the passwords being altered. The two trees were then passed through an audit function to determine the authenticity of the trees. If the root hashes of the two trees differed in any way, then the audit function would return an error message that the tree had been tampered with. In order to parallelize the program, it simply took two OpenMP pragma operations to do so. Then, the parallel implementation was tested using the same methods as for the serial version, and the results were again recorded.

Since there were multiple different running-times for the parallel version depending on the number of processor threads being used, the results were used to construct three line graphs for each of the test cases. The results showed a significant decrease in running-time in the parallel version for at least one, or all of the processor thread options. The

Fig. 13.12 Line graph of the running-time for the parallel implementation using 10,000 passwords

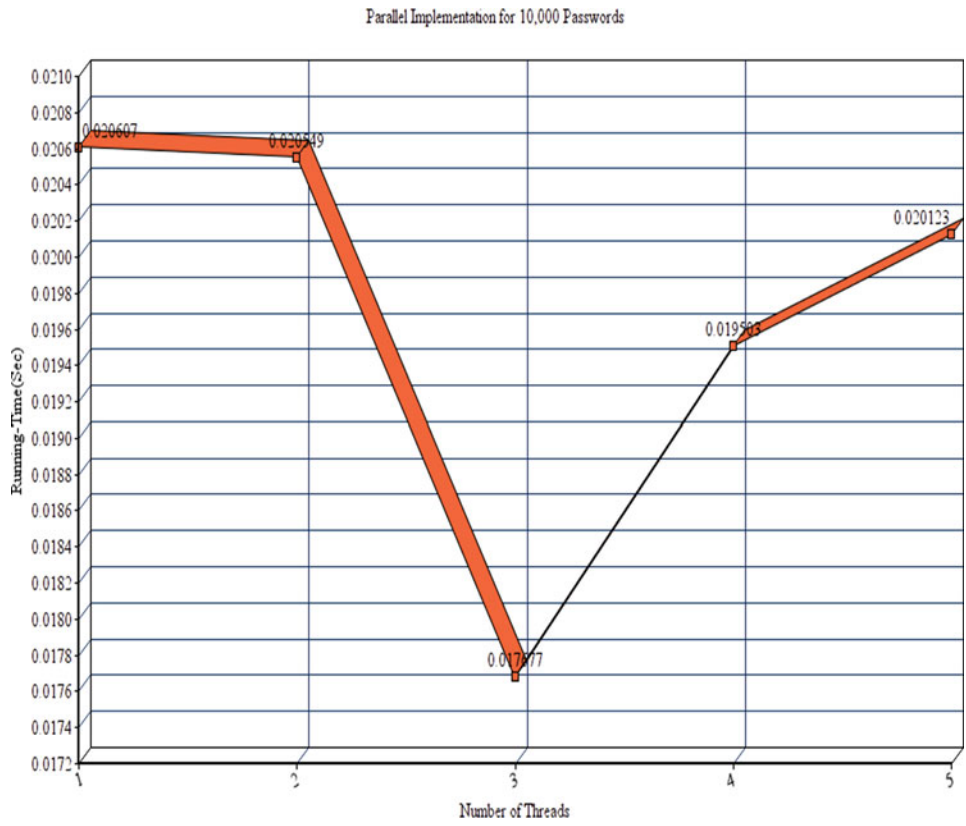
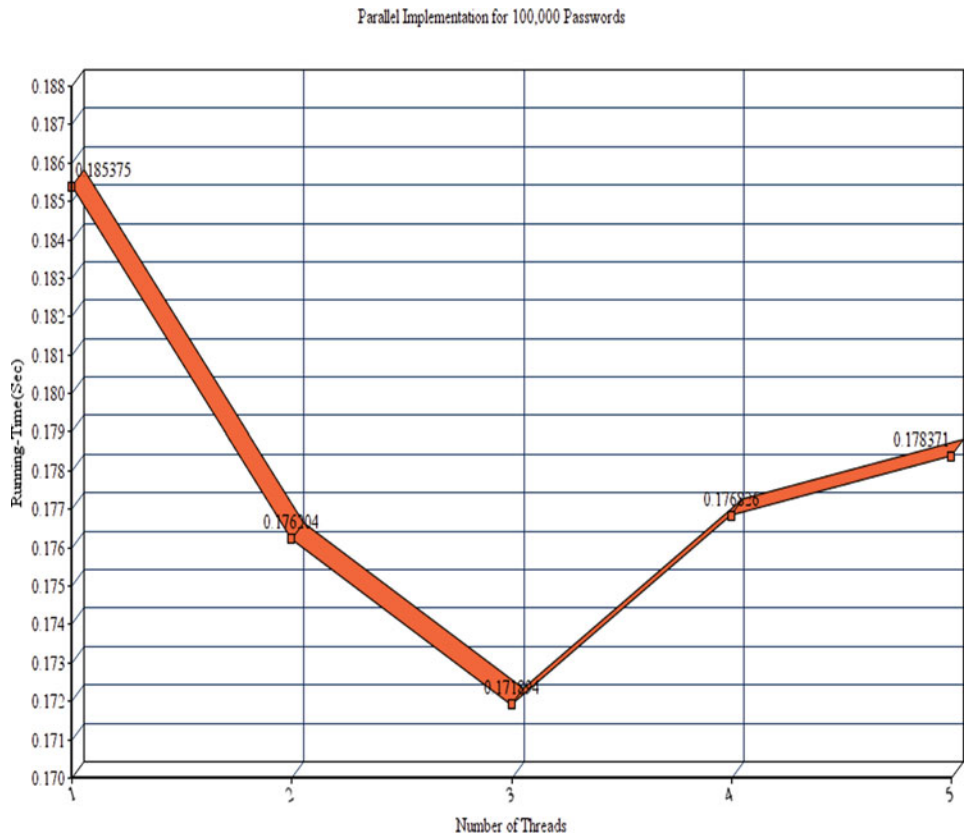


Fig. 13.13 Line graph of the running-time for the parallel implementation using 100,000 passwords



line graphs more clearly represent the data and it shows that the favorable processor thread amount for each of the test cases was three processor threads. For an unclear reason at the moment, the test case of running the program with 10,000 passwords only shows a decrease in running-time when run with three threads, every other processor thread option is slower than the serial version. The best way to remedy this situation may be to use a more effective parallel programming method like MPI or UPC++ to achieve greater results.

13.5.2 Future Work

Some future work to consider for this project could be to develop a parallel version of the same serial implementation as this Merkle tree, but using a more complex MPI version, rather than OpenMP, to see if the running-times can be reduced even further. It would also be beneficial to get better results for the 10,000 password test case as a result of using an MPI implementation, where there is a decrease in running-time from the serial implementation when run on all processors. OpenMP has the ability to parallelize the program by creating threads to run on multiple processors on a single machine, but there are other techniques like that of using UPC++ that can split the workload between different machines and thus, decrease the running time even more. According to the results recorded in this project, the optimal amount of threads to run the Merkle tree program on was three, each running on processors located on the same machine.

The next step for future work could also be to develop a full login application system, where this Merkle tree can be

used in a real-life scenario where a user enters a username and password to log into a system, and if entered incorrectly, will change the root hash. A creative strategy can be used to integrate the root hash of a Merkle tree into the main functionality of the program. At the moment it is unclear what such a program would look like, but the result could be a far more secure system than the ones that are currently accessible to the public. Known exploits for log-in systems can be tested against this system that uses a Merkle tree to store the passwords and then it can be observed if this system is as vulnerable as traditional ones.

References

1. R. Chandra, L. Dagum, D. Kohr, D. Maydan, J. McDonald, R. Menon, *Parallel Programming in OpenMP* (Morgan Kaufmann Publishers, San Francisco, CA, USA, 2001)
2. Wikipedia, Merkle tree, 2020. https://en.wikipedia.org/wiki/Merkle_tree
3. J. Kalidhindi, A. Kazorian, A. Khera, C. Pari, Angela: A sparse, distributed, and highly concurrent merkle tree. Technical report, UC Berkeley, 2018
4. R. Dahlberg, T. Pulls, R. Peeters, Efficient sparse merkle trees: Caching strategies and secure (non-)membership proofs. Cryptology ePrint Archive, Report 2016/683, 2016. <https://eprint.iacr.org/2016/683>
5. K. Atighehchi, R. Rolland, Optimization of tree modes for parallel hash functions: A case study. *IEEE Trans. Comput.* **66**(9), 1585–1598 (2017)
6. B. Gassend, G.E. Suh, D. Clarke, M. van Dijk, S. Devadas, Caches and hash trees for efficient memory integrity verification, in *Proceedings of The Ninth International Symposium on High-Performance Computer Architecture, 2003 (HPCA-9 2003)*, pp. 295–306, 2003
7. A. Miller, M. Hicks, J. Katz, E. Shi, Authenticated data structures, generically. *ACM SIGPLAN Not.* **49**(1), 411–423 (2014)

Marcio Silva Cruz, Ferruccio de Franco Rosa, and Mario Jino

Abstract

Virtual Local Area Network (VLAN) is a technology capable of separating networks into specific domains. Attacks on VLANs could affect computing environments causing service interruptions. These attacks exploit vulnerabilities and operating characteristics of VLANs to gain access to critical information. Conceptual modeling of vulnerabilities and attacks related to VLANs is crucial to enable the construction of systematic methods and techniques for protecting critical infrastructures. Ontologies can contribute in this context, as they are modeling tools that enable the formalization of the main concepts and their relationships, in addition to enabling the creation of semantic rules that can be used by intelligent systems. We present a *quasi*-systematic literature review aiming at describing and classifying studies on ontologies of vulnerabilities and attacks on VLANs. The approach used in this review allowed for the verification and analysis of trends, as well as it uncovers the technological approaches adopted over the past 10 years. The main contributions of this review are: i) a description of the most recent ontologies, taxonomies, techniques and theories, in addition to the contributions and limitations of proposals in the literature; and ii) the identification of gaps in the literature and research challenges. Searches were carried

out in the main scientific knowledge bases in the field of computing. Two hundred sixty-nine articles were found; 19 studies were analyzed according to their approaches, themes and related terms, pointing out contributions and research issues. This article is intended for researchers looking to conceptually model vulnerabilities and attacks on networks.

Keywords

Attack · Network · Ontology · Security · Segmentation · Survey · Taxonomy · Threat · VLAN · Vulnerability

14.1 Introduction

Nowadays, protecting computer networks from cyber attacks is a crucial activity. If we consider the possibility of occurrence of a catastrophic event (e.g., data leakage, denial of service, etc.), to carefully calculate the risks is required. In this context, layer 2 (data link) of the OSI model (Open Systems Interconnection) [1] earns special attention; OSI is a reference model for network protocol projects. VLANs (Virtual Local Area Networks), which are attractive targets for attackers, are in the data link layer [2]. VLAN is a technology capable of separating networks into specific domains. When segmenting the network, workgroups are created to prevent the flow of broadcast frames to devices outside the VLAN, where only the devices of that group, with permission, access resources; this provides good performance of activities, with a defined level of security [3].

Attacks on VLANs could affect upper layers, causing service interruptions and other security problems. These attacks usually exploit the operational characteristics of VLANs to gain access to information. As VLANs are widely used in network environments, a better understanding of their vulner-

M. S. Cruz (✉)

University of Campo LimpoPaulista (UNIFACCAMP), Campo LimpoPaulista, SP, Brazil

F. de Franco Rosa

University of Campo LimpoPaulista (UNIFACCAMP), Campo LimpoPaulista, SP, Brazil

Renato Archer Information Technology Center (CTI), Campinas, SP, Brazil

M. Jino

School of Electrical and Computer Engineering at University of Campinas (UNICAMP), Campinas, SP, Brazil

abilities and all possible attack vectors is required. VLANs must be configured to be non-exposed to known vulnerabilities (e.g., VLAN hopping). For example, in a VoIP (Voice over Internet Protocol) service that uses a VLAN, voice and data traffic are divided into the logical segmentation of the network, minimizing the effects of a possible DoS (Denial of Service) attack [4]. However, the use of softphones on computers, which use the same network interface for data and voice, overrides the application of the VLAN. Thus, a second network interface configured for use of softphones would be necessary, taking into account security recommendations [5].

Identifying and addressing vulnerabilities and their corresponding attacks on VLANs in a systematic and formal manner is essential to keep computing environments secure. However, domains overlap, concepts are ambiguous, terminology is confusing, and important concepts are not defined. Ontologies can contribute in this context [6].

We aim to identify, analyze and classify works that present ontologies or taxonomies of vulnerabilities and attacks on VLANs. The main contributions of this review are: (i) State of the art: description of the most recent ontologies, taxonomies, techniques and theories, in addition to the contributions and limitations of existing proposals; and (ii) Research challenges: identification of gaps in the literature, and current and future research challenges.

The remainder of this article is organized as follows: Section 14.2 presents literature reviews with similar objectives (related work); Sect. 14.3 describes the methodology used to carry out the review and analysis of the articles; Section 14.4 details the results obtained and a categorization of the works; and Sect. 14.5 presents a discussion on the results and the final remarks.

14.2 Review Methodology

The literature review presented in this article was inspired by the methodologies proposed by Kitchenham and De Mendonça et al. [7, 8]. In this study, we aim to answer the following research question: “Which works present or make use of ontological approaches to deal with vulnerabilities and their attack vectors on VLANs?”. Based on the research question, an exploratory search was performed aimed at identifying the search parameters of the literature review, such as search period, scientific databases, keywords and search string, and content area (e.g., title, abstract). The search period (from 2010 to 2020) was defined and the following search string was used (adapted to the syntax of each database): “(*ontology AND vulnerability AND attack AND VLAN AND network*)”.

Two hundred sixty-nine articles were collected, 10 of which were from Springer Link, 12 from IEEE Xplore and 242 from Google Scholar. We have considered all articles returned from the scientific bases, with the exception of

Google Scholar; due to its scope (it indexes other databases), we considered the first 30 articles in order of relevance. The inclusion and exclusion criteria were defined in the exploratory search for an iterative process of reading the papers relevant to the proposed subject. In the first evaluation, we considered title, abstract and keywords. Eleven articles were excluded by analyzing the abstract. Nineteen articles were classified as works with the possibility of adhering to the research subject and were fully evaluated, and a paper from 2004 was found from a quote by another author. During this evaluation, 2 studies were identified as non-adherent; 4 papers are literature reviews related to the subject, and they are described in Sect. 14.3.

14.3 Related Work

During the analysis of the works, 4 systematic reviews were identified. In Appendix I (Table 14.1), we present a summary of the related works. The works were classified according to their objectives and application domains. A description of each literature review is presented below.

Simmonds and Sandilands [9], based their review on standard texts, using well-known concepts, categorizations and methods, e.g., risk analysis using threat profiles, asset-based and vulnerability profiles. Network security services, threat analysis, vulnerabilities and failure modes were also considered. The review is used to build a framework, which is used to define an extensible ontology of network security attacks.

Bijani and Robertson [10] present and classify the main attacks on open MASs (Multi-agent Systems). In this literature review, the authors search and analyze the various security techniques and categorize them as prevention and detection approaches. Additionally, the authors suggest which security technique is an appropriate countermeasure for which classes of attack.

Luh et al. [11] argue that there is a lack of focus on advanced persistent threats (APT) (sophisticated multi-stage attacks) and suggest that developing appropriate methodologies to act in this domain is a research challenge. In this context, the authors present a detailed literature review scheme, in addition to a model for categorizing articles. The selected articles were analyzed and evaluated according to the Kitchenham guidelines [7]. They combine new insights and the status quo of current research with the concept of an ideal systemic approach, capable of semantically processing and evaluating information on different aspects. The papers presented contribute to the analysis or detection of targeted attacks.

Singh et al. [12], present a study on the contributions that analyze and detect advanced persistent threats (APTs). The exploratory research covered various aspects, e.g., the

exploitation of web infrastructure and communication protocols. The authors present a comprehensive literature assessment scheme that classifies and provides countermeasures to APT attacks.

Our literature review differs from other reviews in the following aspects: (1) the focus of our review is on ontologies and taxonomies of vulnerabilities and attacks on VLANs, highlighting attacks to the data link layer; (2) existing reviews focus on aspects of vulnerabilities, attacks and security for solutions that aim at the security status of networks in general; they do not cover, in a broader discussion, the risks and problems arising from the use of segmented networks; and, (3) the studies do not address vulnerability analyzes of specific-layer protocols.

14.4 Ontologies of Vulnerabilities and Attacks on VLANs

We present an analysis of 14 selected studies, which present solutions based on ontologies for analyzing vulnerabilities and attacks on computer systems. Subsection 14.4.1 presents works that make use of ontologies; Subsection 14.4.2 presents works that propose ontologies. Appendix II (Table 14.2) presents a summary of the analyzed works.

14.4.1 Works that Use Ontologies

Bhandari and Gujral [13] proposed an ontology to infer the impact of network security incidents. Vulnerability, Network and Attack are the main classes of the proposed ontology. According to the authors, Computer Network is a dynamic entity whose status changes constantly, e.g., with the introduction of new services, installation of a new operating system, addition of new hardware components, creation of new user features, etc.; it could insert new vulnerabilities in the computing environment. Various security mechanisms are used in the network, but do not provide the complete security view of the entire network. A framework, which uses this ontology as a knowledge base, was proposed to provide situational awareness of network security.

Shenbagam and Salini [14] propose an ontology-based approach to defend against attacks on application vulnerabilities. The proposed attack prediction system is based on an ontology of attacks to web applications.

Si et al. [15] propose a method of fusing elements of the network security situation based on ontology; the fusion model contains the network environment, network vulnerability, network attack, network security incident, and the sensor as the main class.

Choi et al. [16] analyze the source code of an APT (Advanced Persistent Threat) to propose a method for detecting

these attacks. The proposed method is based on an attack behavior ontology. An intelligent APT attack is used to define rules of inference about the attack behavior.

Krau and Thomalla [17] present an ontology-based approach for detecting cyber attacks on SCADA (Supervisory Control and Data Acquisition) systems. System logs provide events that intrusion detection systems (IDS) could recognize as suspicious and that could be part of an attack. The proposed model uses databases of known vulnerabilities to identify ongoing attacks.

Xu et al. [18] propose a network security situation recognition model for IoT (Internet of Things) devices. The proposal uses a situation reasoning method based on ontology and user-defined semantic rules. According to the authors, ontologies can provide a unified and formalized description to solve the problem of semantic heterogeneity in the IoT security domain.

14.4.2 Works that Propose Ontologies

Gao et al. [19] classify attacks in a security assessment taxonomy and present an ontology-based framework for assessing the security of systems and computer networks. The authors describe the use of ontology in assessing security and the method for assessing the effect of attacks on the system when it is under attack. The proposed taxonomy consists of 5 dimensions, namely: attack impact, attack vector, attack target, vulnerability and defense. The concepts are analyzed, related to each other, and formalized to generate an ontology.

Karande and Gupta [20] propose an intrusion detection system (IDS) based on an ontology of Web application security. The proposed system aims to obtain context information through links and scripts; the ontology model establishes a semantic relationship between attacks and networks. The proposed IDS ontological model allows to detect attacks through specific protocols and identifying malicious scripts, in addition to identifying the types of attacks and vulnerabilities. According to the authors, the proposed ontology is recommended to describe security concepts of Web services.

Kshirsagar et al. [21] propose an ontology for detecting attacks that exploit the HTTP response splitting vulnerability. The proposed ontology allows to generate semantic rules.

Chavan and Tamane [22] present an ontology for detecting attacks on cloud-based web services. Based on cloud architectures, security rules for attacks on web applications can be specified. In this work, other approaches are also discussed, e.g., detection of malicious traffic over the Internet.

Mohsin and Anwar [23] propose an ontological framework for IoT to protect against APTs. The approach involves understanding attack patterns and vulnerabilities, and aligning them with network semantics to assess their feasibility in IoT systems. Ontologies of cyber threats intelligence stan-

dards are extended with new concepts and aligned with a new IoT ontology.

Falodiya and Das [24] present an ontology for attack graphs aimed at analyzing security vulnerabilities in corporate networks. According to the authors, attack graphics support the modeling of security vulnerabilities, as well as the identification of possible paths in a corporate network that could be used by an attacker to exploit network vulnerabilities.

Choi and Choi [25] propose a security context ontology based on the analysis of security vulnerabilities of a power system in an energy IoT-Cloud environment. The ontology allows the definition of rules for inference of the security context of critical infrastructure.

Zhu et al. [26] propose an ontology of vulnerabilities, based on public information security databases. The main purpose is to standardize and describe information about known vulnerabilities.

14.5 Discussion and Final Remarks

Maintaining information security and privacy in a cloud environment is a critical issue [22]. The increasing use of Web applications leads to a large number of threats and vulnerabilities; 81% of hacker attacks target Web applications, which pose a major threat to the security of online banking, e-commerce, etc. [21]. Security of Web applications is the main concern in the context of e-business and information sharing; 75% of attacks are performed in the application layer and almost 90% of web applications have some vulnerability [14].

Studies on the risks and vulnerabilities arising from the use of VLANs in computing environments are rare. We can infer from our survey the imperative need to protect segmented networks, which are often neglected due to the lack of knowledge about their vulnerabilities and attack vectors. We identify solutions based on ontologies to support methods and models, analyze vulnerabilities, and to detect attacks on computer systems. In the context of conceptual modeling aimed at protecting VLANs, ontologies could be used to formalize and relate important concepts, map vulnerabilities and known attacks, describe protection processes, generate rules for inference, among other applications.

We presented a *quasi*-systematic literature review, by describing and classifying studies on ontologies of vulnerabilities and attacks on VLANs. Our review approach allowed us to verify and analyze trends, and identify research challenges as well. From 269 articles initially selected, 18 works were carefully analyzed, classified and summarized in order to represent the state-of-the-art. The techniques and theories used were presented, positive aspects and limitations of the studies were discussed, and gaps in the literature and research challenges were unveiled.

In addition to presenting a comprehensive literature review on ontologies of vulnerability and attacks on VLANs, this work also contributes to proposals for methods, processes and standards based on ontology aimed at mitigating attacks on segmented networks through systematic and formal techniques.

Appendix I

Table 14.1 Summary of related work

Authors	Objective						Application domain			
	Ou	Op	Mt	Md	F	T	1	2	3	4
Simmonds and Sandilands [9]		X		X		X	X	X	X	X
Bijani and Robertson [10]				X		X		X	X	X
Luh et al. [11]				X		X		X		X
Singh et al. [12]				X		X		X	X	
<i>Our work</i>	X	X	X	X		X	X	X	X	X

Objective: Ontology – using of (Ou); Ontology – proposal of (Op); Method (Mt); Model (Md); Framework (F); Taxonomy (T)

Application Domain: (1) Vulnerabilities; (2) Attacks or Detection; (3) Mitigation or Defense; (4) Status – Network or Security

Appendix II

Table 14.2 Summary of analyzed work

Authors	Objective						Application Domain			
	Ou	Op	Mt	Md	F	T	1	2	3	4
Bhandari and Gujral [13]	X			X		X	X	X	X	X
Shenbagam and Salini [14]	X		X				X	X		
Si et al. [15]	X		X							X
Choi et al. [16]	X		X					X	X	
Krauß and Thomalla [17]	X		X					X	X	
Xu et al. [18]	X			X						X
Gao et al. [19]	X	X		X		X	X	X	X	X
Karande and Gupta [20]		X		X			X	X	X	
Kshirsagar et al. [21]		X		X				X		
Chavan and Tamane [22]		X		X				X		X
Mohsin and Anwar [23]		X		X			X	X	X	X
Falodiya and Das [24]		X	X				X	X	X	
Choi and Choi [25]		X		X			X		X	X
Zhu et al. [26]		X		X			X			

Objective: Ontology – using of (Ou); Ontology – proposal of (Op); Method (Mt); Model (Md); Framework (F); Taxonomy (T)

Application Domain: (1) Vulnerabilities; (2) Attacks or Detection; (3) Mitigation or Defense; (4) Status – Network or Security

References

1. A.S. Tanenbaum, D. Wetherall, *Computer Networks*, 5th Edition, USA (2011)
2. S. Convery, *Network Security Architectures – Expert Guidance on Designing Secure*, First prin. (Cisco Press, Indianapolis – USA, 2004)
3. O. Soares Barros, Segurança de redes locais com a implementação de VLANs O caso da Universidade Jean Piaget de Cabo Verde. p. 67, 2006, [Online]. Available: <http://hdl.handle.net/10961/4220%0A> (in Portuguese)
4. P. Thermos, A. Takanen, *Securing Voip Networks: Threats, Vulnerabilities and Countermeasures*, 1st edn. (Addison-Wesley Professional, 2007)
5. P. Thermos, A. Takanen, *Securing Voip Networks: Threats, Vulnerabilities and Countermeasures*, 1st Edition. (Addison-Wesley Professional, 2007) Publisher : Addison-Wesley Professional; 1st Edition (August 11, 2007), All rights reserved. Printed in the United States of America. ISBN-13: 978- 0-321-43734-1 ISBN-10: 0-321-43734-9
6. F. de Franco Rosa, M. Jino, R. Bonacin, Towards an ontology of security assessment: a core model proposal, in *Advances in Intelligent Systems and Computing*, vol. 738, (2018), pp. 75–80. https://doi.org/10.1007/978-3-319-77028-4_12
7. B. Kitchenham, Procedures for performing systematic reviews. Keele, UK, Keele Univ. **33**(TR/SE-0401), 28 (2004). [10.1.1.122.3308](https://doi.org/10.1.1.122.3308)
8. R.R. de Mendonça, F. de Franco Rosa, A.C. Theophilo Costa, R. Bonacin, M. Jino, OntoCexp: a proposal for conceptual formalization of criminal expressions, in *16th International Conference on Information Technology-New Generations (ITNG 2019)*, no. Itng, 2019, pp. 43–48
9. A. Simmonds, P. Sandilands, L. Van Ekert, “An ontology for network security attacks,” in *Asian Applied Computing Conference*, 2004, pp. 317–323
10. S. Bijani, D. Robertson, A review of attacks and security approaches in open multi-agent systems. *Artif. Intell. Rev.* **42**(4), 607–636 (2014). <https://doi.org/10.1007/s10462-012-9343-1>
11. R. Luh, S. Marschalek, M. Kaiser, H. Janicke, S. Schrittwieser, Semantics-aware detection of targeted attacks: a survey. *J. Comput. Virol. Hacking Tech.* **13**(1), 47–85 (2017). <https://doi.org/10.1007/s11416-016-0273-3>
12. S. Singh, P.K. Sharma, S.Y. Moon, D. Moon, J.H. Park, A comprehensive study on APT attacks and countermeasures for future networks and communications: challenges and solutions. *J. Supercomput.* **75**(8), 4543–4574 (2019). <https://doi.org/10.1007/s11227-016-1850-4>
13. P. Bhandari, M.S. Gujral, Ontology based approach for perception of network security state, *2014 Recent Adv. Eng. Comput. Sci. RA ECS 2014*, pp. 6–8, 2014, <https://doi.org/10.1109/RAECS.2014.6799584>
14. J. Shenbagam, P. Salini, Vulnerability Ontology for web applications to predict and classify attacks, *2014 Int. Conf. Electron. Commun. Comput. Eng. ICECCE 2014*, pp. 268–272, 2014, <https://doi.org/10.1109/ICECCE.2014.7086625>
15. C. Si, H. Zhang, Y. Wang, J. Liu, Network security situation elements fusion method based on ontology, *Proc. – 2014 7th Int. Symp. Comput. Intell. Des. Isc. 2014*, vol. 2, pp. 272–275, 2015, <https://doi.org/10.1109/ISCID.2014.132>
16. J. Choi, C. Choi, H.M. Lynn, P. Kim, Ontology based APT attack behavior analysis in cloud computing, *Proc. – 2015 10th Int. Conf. Broadband Wirel. Comput. Commun. Appl. BWCCA 2015*, pp. 375–379, 2015, <https://doi.org/10.1109/BWCCA.2015.69>
17. D. Krauß, C. Thomalla, Ontology-based detection of cyber-attacks to SCADA-systems in critical infrastructures,” *2016 6th Int. Conf. Digit. Inf. Commun. Technol. Its Appl. DICTAP 2016*, pp. 70–73, 2016, <https://doi.org/10.1109/DICTAP.2016.7544003>
18. G. Xu, Y. Cao, Y. Ren, X. Li, Z. Feng, Network security situation awareness based on semantic ontology and user-defined rules for internet of things. *IEEE Access* **5**, 21046–21056 (2017). <https://doi.org/10.1109/ACCESS.2017.2734681>
19. J.B. Gao, B.W. Zhang, X.H. Chen, Z. Luo, Ontology-based model of network and computer attacks for security assessment. *J. Shanghai Jiaotong Univ.* **18**(5), 554–562 (2013). <https://doi.org/10.1007/s12204-013-1439-5>
20. H.A. Karande, S.S. Gupta, Ontology based intrusion detection system for web application security, pp. 228–232, 2015, <https://doi.org/10.1109/icc.2015.44>
21. D. Kshirsagar, S. Kumar, L. Purohit, Exploring usage of ontology for HTTP response splitting attack, *Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol. NGCT 2015*, no. September, pp. 437–440, 2015, <https://doi.org/10.1109/NGCT.2015.7375156>
22. S.M. Chavan, S.C. Tamane, Study and design of ontology for cloud based web services attacks: a survey, *Proc. – Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun. ICGTSPICC 2016*, pp. 24–29, 2016, <https://doi.org/10.1109/ICGTSPICC.2016.7955263>
23. M. Mohsin, Z. Anwar, Where to kill the cyber kill-chain: an ontology-driven framework for IoT security analytics, *Proc. – 14th Int. Conf. Front. Inf. Technol. FIT 2016*, pp. 23–28, 2016, <https://doi.org/10.1109/FIT.2016.013>
24. K. Falodiya, M.L. Das, Security vulnerability analysis using ontology-based attack graphs, *2017 14th IEEE India Counc. Int. Conf. INDICON 2017*, pp. 1–5, 2018, <https://doi.org/10.1109/INDICON.2017.8488002>
25. C. Choi, J. Choi, Ontology-based security context reasoning for power IoT-cloud security service. *IEEE Access* **7**, 110510–110517 (2019). <https://doi.org/10.1109/access.2019.2933859>
26. L. Zhu, Z. Zhang, G. Xia, C. Jiang, Research on vulnerability ontology model, *Proc. 2019 IEEE 8th Jt. Int. Inf. Technol. Artif. Intell. Conf. ITAIC 2019*, no. Itaic, pp. 657–661, 2019, <https://doi.org/10.1109/ITAIC.2019.8785783>

Daniel Alarcón-Narváez and Fausto A. Jacques García

Abstract

In symmetric encryption, the algorithm and secret key determine the security factor. This paper presents the idea to create a multiple (7 times) and block-separated encryption algorithm. To achieve this, we will use the Hill Cipher and Gauss-Jacques methods. In addition to the above, our most significant contribution will be to obtain large secret keys, which will allow us to obtain as a possible result a meaningful approximation to the Shannon perfect secrecy, as well as the reduction of computational complexity and the verification of security through anti-bot mechanisms such as code breakers.

Keywords

AES · Block cipher · Ciphertext · Cryptography · Gauss-Jacques · Hill cipher · Modular inverse matrix · Plaintext · Secret key · Symmetric encryption

15.1 Introduction

Symmetric encryption is a form of cryptosystem in which encryption and decryption are performed using the same key. In other words, those that require both parties to use the same secret key. Algorithms that use a shared key are known as symmetric algorithms. Figure 15.1 illustrates the basic encryption of symmetric key. It is also known as conventional encryption. The two types of attacks on an encryption algorithm are crypto-analysis, based on the encryption algorithm's properties and brute-force, which involves testing all

possible keys. Any symmetric encryption scheme contains five main parts:

- Plaintext: This is the message or data to be encrypted and it is the input of the algorithm.
- Encryption algorithm: It performs substitutions or transformations on the plaintext.
- Secret key: It is also an input to the algorithm. That depending on the key that is provided, will be the output that we will obtain.
- Ciphertext: This is the message or encoded data obtained from the plaintext and the secret key as output.
- Decryption Algorithm: This is essentially the encryption algorithm that runs in reverse. It takes the ciphertext and the secret key and produces the original plaintext.

In symmetric encryption, they use two techniques: substitution or transposition. Substitution techniques map the elements of the plaintext to elements of the ciphertext. The transposition techniques systematically transpose the positions of the elements of the plaintext. For our work, we will focus on a substitution technique called Hill Cipher. Below we listed the most common techniques in the literature [1].

- Caesar Cipher
- Monoalphabetic Ciphers
- Playfair Cipher
- Hill Cipher
- Polyalphabetic Ciphers
- One-Time Pad
- Advanced Encryption Standard (AES)

The advanced encryption standard is one of the most popular symmetric ciphers and, therefore, the most used. We mention this because our proposed algorithm pretends to be the same or more secure, with the difference also that our algorithm will require a less computational cost.

D. Alarcón-Narváez (✉) · F. A. Jacques García
 Computer Science School, Autonomous University of Queretaro,
 Queretaro, Mexico
 e-mail: dalarcon15@alumnos.uaq.mx; jacques@uaq.edu.mx

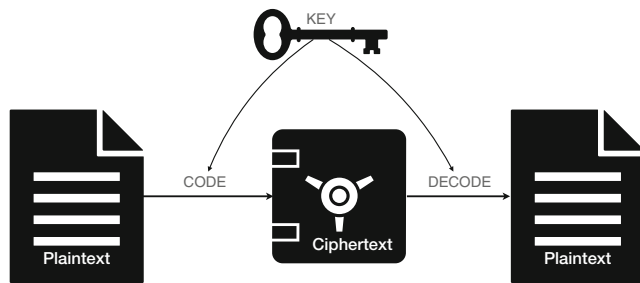


Fig. 15.1 Symmetric key encryption (Source: Own design)

The US National Institute of Standards and Technology, in October 2000, selected the block cipher Rijndael as the Advanced Encryption Standard (AES) [2]. Various tests have been carried out and have been given many applications to the Rijndael. For example, Ferguson et al. [3], assessed the safety of the Rijndael and described various attacks as well as unanticipated properties of the cipher. Although they only analyzed the 120-bit version. Therefore, the other variants were analyzed separately due to their differences between them. Also, in that article, they introduced a technique that they called partial summation, which reduces the work factor of the attack.

Nawal and Osama [4], compared three-block cipher algorithms, RC6, MRC6, and Rijndael. They encrypted different types of images with each one of them. They considered different measurement factors between the original image and the encrypted image, based on measuring the maximum deviation, the correlation coefficient, the difference of the pixel, the encryption time, and the throughput. They obtained as a result that the Rijndael algorithm is the best.

Lee et al. [5], proposed a Rijndael-based block encryption/decryption method, which includes an operating unit that performs encryption/decryption. The encryption apparatus is mounted on a mobile terminal as a cell phone and a PDA or smart card, requiring a small, high-speed encryption processor, the above to encrypt and decrypt essential data required high-speed and security. As a result, they found that the encryption apparatus can reduce the time required for encryption/decryption of other block codes and the apparatus's size.

Srinivas and Akramuddin [6], presented the hardware implementation of the AES Rijndael encryption and decryption algorithm using Xilinx Virtex-7 FPGA. The hardware design approach was based entirely on pre-computed lookup tables (LUTs), which resulted in a less complicated architecture, as well as high performance and low latency. They used three different formats in AES, 128, 192, and 256. The encryption and decryption blocks of the three formats were efficiently designed using Verilog-HDL and synthesized on the Virtex-7 XC7VX690T chip with the help of the Xilinx ISE Design Suite-14.7 tool. As a result, they found that the architecture

they proposed had good efficiency in latency, throughput, speed/delay, area, and power.

As we mentioned above, in our work, we will focus on the Hill Cipher method. Different studies have been carried out around Hill Cipher, for example.

Paragas et al. [7], used the Hill Cipher method, which is vulnerable to attacks against a known plaintext. To solve this and other limitations, they used plaintext 128-bit block encryption using multiple encryption processing rounds, encryption block chaining, and hexadecimal substitution box. This modified approach offered better data protection and overcame the drawbacks of the original Hill Cipher algorithm.

The test results showed that the ciphertext protection has an avalanche effect score of 55.34%. In this study, 3 phases were applied, the generation of the key matrix where they used SHA256 to find a 32-bit key and thus be able to divide it into 4 blocks of 8 bits. The plaintext encryption process were divided into 16 characters per block; the first row was rounded left and moved to the last row of the plaintext matrix with four character spaces. For the rest of the block cipher method, it was done four times. The result goes through logical XOR operations using CBC1 before moving to the hex substitution box. The decryption process was repeated in reverse.

Man, K. et al. [8], talked about the importance of selecting a correct key matrix because many times, it is not possible to find the inverse of a matrix. So they used any of the classic ciphers, like Playfair Cipher, ADFGVX Cipher, etc. In the article, the key matrix was not generated randomly but was generated using predetermined sequential advance and permutation procedures.

Khalaf et al. [9], studied the safety problem and presented a triple Hill Cipher algorithm. It was implemented in FPGA to encrypt any binary data such as images, audio, video, etc. That algorithm used three stages of the modified Hill Cipher to make the algorithm more robust. Each stage is considered a block cipher with a block length of 128 bits and a key length of 256 bits. The message to be encrypted is processed by this block encryption in three stages to increase security. In other words, the plaintext is encrypted using eight rounds with eight different keys three times; we can say that they used 24 different keys, making it difficult to attack the plaintext. The key matrices are taken from the random number generator.

Saednia [10] explained in his article how the Hill Cipher method is well known for being insecure, saying that the weakness comes from the fact that the cipher only provides linear transformations; that is, they are linearly dependent. Although in this work it only focuses on making an insecure method safe by exchanging a matrix only once and thus reducing the calculation of inverse matrices and performing it with less complicated operations. So efficiency and speed are not in the author's interest.

Jangid et al. [11] presented a new proposal for the Hill-Cipher method, using DNA cryptography and TFHill Cipher to solve the attack problems that another method has. This algorithm was used to encrypt images where it makes a variation of the colors to gray scales to later convert it into binary 1 for white and 0 for black. Then they convert it into DNA and later into amino acids. As a result, they obtained a higher entropy value and a lower correlation, and a more uniform histogram. The proposed cryptosystem thwarts chosen-ciphertext, known plaintext, and chosen-plaintext attacks.

One of the disadvantages encounters in Hill Cipher is that it uses the inverse of the key matrix, wherein on some occasions, the inverse matrix cannot be found. It also uses the determinant of a matrix, which, for the cases of huge matrices, it is complicated to calculate its value [12, 13].

Inverse matrices are quite common in Cryptography, so many articles have been published to improve its implementation. For example, Asad et al. [14] implemented in a 128-bit inverse modular unit programmable device using the Extended Euclidean Algorithm. The design was described by the VHDL language, the simulation was done by Mod-elSim and the synthesis by Quartus II tools. The implementation results showed a comparative cost of factors that can improve the performance of many applications that include investment operations in their calculations.

W. Bos [15] showed how to modify a Kaliski algorithm to calculate the classical modular inversion and the Montgomery inversion in constant time. It is mentioned that an effective countermeasure to protect an implementation is to guarantee a constant execution time (in the worst case). Consistent execution time is an essential countermeasure against simple power analysis attacks with applications in public-key cryptography. He showed that in popular ARM architecture, this approach outperforms current approaches based on Fermat's theorem when using generic prime modules. In the context where prime numbers can be used, the modular inversion approach based on Fermat's theorem could be more efficient.

Phiamphu and Saha [16] proposed an improved Extended Euclidean Algorithm (EEA) architecture to find the Modular Multiplicative Inverse (MMI) and the Jacobi symbol. The method was implemented in the HDL hardware description language. The results were compared with the existing extended Euclidean algorithm (EEA). It was showing superiority in the results in all aspects when compared with existing methods.

Martin Seysen [17] presented a new algorithm to calculate modular inversion, replacing the extended Euclidean algorithm with a standard Euclidean algorithm. This algorithm works with integers, which are twice the length of the module. Obtaining twice the speed in the applications that it is given in cryptology.

Bajard et al. [18], argued the use of the Residue Number System (RNS), its properties, and its applications in cryptology. Therefore, the resulting arithmetic is resistant to attacks. RNS is suitable for RSA but not really for ECC. Thus, in their work, they analyzed modular investment characteristics in RNS over GF (P). Proposing an RNS extended Euclidean algorithm that uses a quotient approximation module.

To calculate the inverse matrix in our work, we will use the Gauss-Jacques method (Eq. 15.1), giving the inverse matrix more easily and quickly. The Gauss-Jacques method's main objective is to obtain modular inverse matrices of size $n \times n$. This method reduces the computational cost compared to that used by the Hill-Cipher method. For example, Hill cipher calculates the inverse matrix using the determinant of a matrix with a computational cost of $\mathcal{O}(n!)$. Unlike the Gauss-Jacques calculates the inverse matrix with a computational cost of $\mathcal{O}(\phi^n)$.

$$K_m^{-1} = \left\{ \sum_{i=1}^s \sum_{j=1}^s [-k_{j(i+1)} + (k_{ij}e(k_{ii}, m) \bmod m)] \bmod m \right\} \quad (15.1)$$

With these two methods, the Hill Cipher and the Gauss-Jacques, we hope to achieve 3 points: a multiple encryption model (7 times), separated by blocks, and with a Key matrix at the most countable. Figure 15.2 shows an example of multiple encryption per block and Fig. 15.3 shows an example of multiple decryption per block. Therefore, the work's general objective will be to create a new symmetric encryption model capable of achieving perfect secrecy. To complete the general-purpose, we will have three main goals:

1. Create a block cipher method for the Hill Cipher.
2. Create a 7 times multiple encryption method.
3. Generate Key Matrix to the most countable.

We organized this paper as follows: In Sect. 15.2, we presented the two methods to use in a general way. Section 15.3 describes part of the proposed method to achieve our goal. Finally, in Sect. 15.4, we summarized the results we hope to obtain.

15.2 Methods

15.2.1 Hill Cipher

Hill Cipher, developed by mathematician Lester Hill in 1929. The main focus of this method is the manipulation of matrices. The formulas to obtain the Ciphertext and the original text from the inverse Key Matrix are as follows [12, 13]:

$$C = (K * P) \bmod 26 \quad (15.2)$$

Fig. 15.2 Example of multiple encryption per block (Source: Own design)

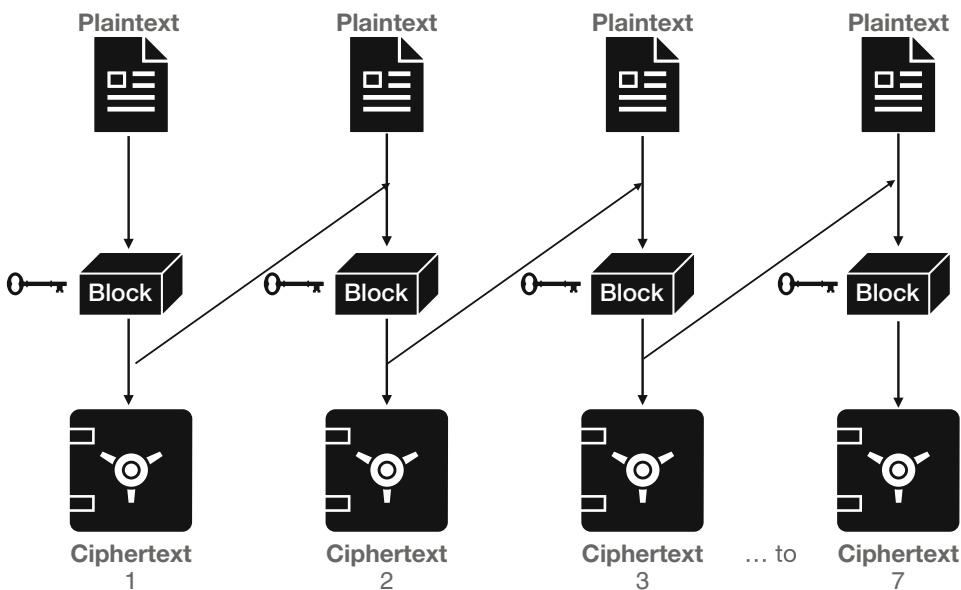
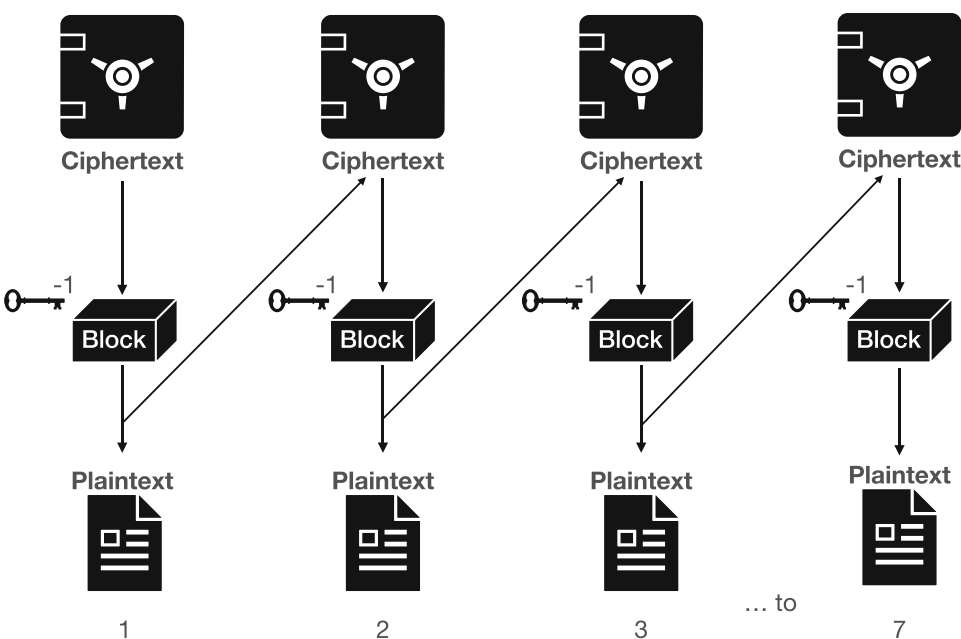


Fig. 15.3 Example of multiple decryption per block (Source: Own design)



where:

- C = Ciphertext
- K = Key
- P = Plaintext

$$P = (K^{-1} * C) \text{mod} 26 \tag{15.3}$$

Where:

$$K^{-1} = \frac{1}{|K|} \text{adj} K \tag{15.4}$$

15.2.2 Gauss-Jacques

In this method, the modular inverse matrix is calculated using the echelon and reduced form by rows, the Euclidean modulus and the extended Euclidean algorithm, plus the product between two matrices as a check. The pseudo-code of the technique is shown in the following list [19, 20]:

1. The size of the Key is declared
2. The Key matrix is created with random numbers
3. Choose a prime module
4. Write the identity matrix next to the Key matrix (right side)

5. Find a number that multiplied by position a_{11} of the matrix and the module result one.
6. Then, The formula is applied to the entire line
7. The pivot is used to reduce the elements of its column to zero and apply to the whole row
8. It is done again from step 5 to position a_{22} , stepping up to have the identity matrix on the left side
9. The matrix on the right side is the inverse matrix

C = Ciphertext

$\text{mod } m$ = The modulus of a prime number

$K_{m_i}^{-1}$ = The inverse Matrix

15.3 Methodology

In this work, we will use 2 methodologies: Design-based Research Methodology and Experimental Methodology. Design-based Research Methodology is a type of research oriented towards educational innovation whose fundamental characteristic consists in the introduction of a new element to transform a situation. This type of research tries to respond to problems detected in the educational reality by resorting to scientific theories or available models in order to propose possible solutions to these problems [21]. Experimental Methodology implies the observation, manipulation, recording of the variables that affect an object of study [22].

Equations 15.5 and 15.6 show the formulas that we propose to perform block encryption and gradually obtain either a plaintext or a ciphertext for each block. Equations 15.7 and 15.8 show the formulas to perform multiple encryption and decryption for 7 times. It is necessary to mention that when we obtain the result of the operations in the blocks, we will carry out the multiple “refinement” 7 times. In other words, we will replace the result of Eq. 15.5 in Eq. 15.7, just as the result of Eq. 15.6 in Eq. 15.8. We hope that with the application of multiple encryption and block encryption, as well as with the most countable matrix, we will achieve a high security factor. Part of the proposal that it be done 7 times is to observe the effect on the variables. This proposal is based on a refining analogy as gold that is done 7 times.

15.3.1 Block

$$\bigcup_{i=1}^n \{(K_i P_i) \text{ mod } m\} = C \quad (15.5)$$

$$\bigcup_{j=1}^n \{(K_{m_i}^{-1} C_i) \text{ mod } m\} = P \quad (15.6)$$

Every formula in this subsection can be defined by using six tuples where:

\bigcup = Every block created from 1 to n

K = Key Matrix

P = Plaintext

15.3.2 7-HAJ (Multiple)

$$\sum_{k=1}^7 \{C_k\} = 7C \quad (15.7)$$

$$\sum_{l=1}^7 \{P_l\} = 7P \quad (15.8)$$

Every formula in this subsection can be defined by using three tuples where:

\sum = The sum of the multiple encryption or decryption from 1 to 7

P = Plaintext

C = Ciphertext

These formulas are expected to obtain a model as efficient as safe with an immediate IoT application. We hope we can improve computational complexity and security.

15.4 Expected Results

In this project, we propose a new algorithm to encrypt and be used in the computational area from a Symmetric point of view. Hence, we expect to create this new symmetric encryption model capable of achieving perfect secrecy. So we hope to create libraries for some particular software, and we hope to register it as a new crypto idea. Therefore, it is also expected to make more publications with the results obtained.

We are open to any contribution or suggestion to enrich our work.

References

1. W. Stallings, *Cryptography and Network Security, 4/E* (Pearson Education India, 2006)
2. J. Daemen, V. Rijmen, *The Design of Rijndael: The Advanced Encryption Standard (AES)* (Springer Nature, 2020)
3. N. Ferguson, J. Kelsey, S. Lucks, B. Schneier, M. Stay, D. Wagner, D. Whiting, Improved cryptanalysis of rijndael, in *International Workshop on Fast Software Encryption* (Springer, 2000), pp. 213–230
4. N.F. El Fishawy, O.M.A. Zaid, Quality of encryption measurement of bitmap images with rc6, mrc6, and rijndael block cipher algorithms. *IJ Network Security* 5(3), 241–251 (2007)
5. Y.K. Lee, Y.S. Park, Y.S. Kim, S.W. Lee, S.I. Jun, Rijndael block cipher apparatus and encryption/decryption method thereof, Mar. 30 2010, US Patent 7,688,974
6. N.S. Srinivas, M. Akramuddin, Fpga based hardware implementation of aes rijndael algorithm for encryption and decryption,

- in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (IEEE, 2016), pp. 1769–1776
7. J.R. Paragas, A.M. Sison, R. Medina, An improved hill cipher algorithm using cbc and hexadecimal s-box, in *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)* (IEEE, 2019), pp. 77–81
 8. R. Mahendran, K. Mani, Generation of key matrix for hill cipher encryption using classical cipher, in *2017 World Congress on Computing and Communication Technologies (WCCCT)* (IEEE, 2017), pp. 51–54
 9. A.A. Khalaf, M.S. Abd El-karim, H.F. Hamed, Proposed triple hill cipher algorithm for increasing the security level of encrypted binary data and its implementation using fpga, in *2015 17th International Conference on Advanced Communication Technology (ICTACT)* (IEEE, 2015), pp. 454–459
 10. S. Saeednia, How to make the hill cipher secure. *Cryptologia* **24**(4), 353–360 (2000)
 11. R.K. Jangid, N. Mohmmad, A. Didel, S. Taterh, Hybrid approach of image encryption using dna cryptography and tf hill cipher algorithm, in *2014 International Conference on Communication and Signal Processing* (IEEE, 2014), pp. 934–938
 12. B. Acharya, S.K. Panigrahy, S.K. Patra, G. Panda, Image encryption using advanced hill cipher algorithm. *Int. J. Recent Trends Eng.* **1**(1), 663–667 (2009)
 13. B. Acharya, G.S. Rath, S.K. Patra, S.K. Panigrahy, Novel methods of generating self-invertible matrix for hill cipher algorithm. *Int. J. Secur.* **1**, 14–21 (2007)
 14. M.M. Asad, I. Marouf, Q.A. Al-Haija, A. AlShuaibi, Performance analysis of 128-bit modular inverse based extended euclidean using altera fpga kit. *Procedia Comput. Sci.* **160**, 543–548 (2019)
 15. J.W. Bos, Constant time modular inversion. *J. Cryptographic Eng.* **4**(4), 275–281 (2014)
 16. D. Phiamphu, P. Saha, Redesigned the architecture of extended-euclidean algorithm for modular multiplicative inverse and jacobi symbol, in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)* (IEEE, 2018), pp. 1345–1349
 17. M. Seysen, Using an rsa accelerator for modular inversion, in *International Workshop on Cryptographic Hardware and Embedded Systems* (Springer, 2005), pp. 226–236
 18. J.C. Bajard, N. Meloni, T. Plantard, Study of modular inversion in rns, in *Advanced Signal Processing Algorithms, Architectures, and Implementations XV*, vol. 5910. International Society for Optics and Photonics, 2005, p. 59100T
 19. F.A.J. García, El método gauss-jacques propuesto para la obtención de matrices inversas modulares de tamaño variable sin límite teórico. *Digital Ciencia@UAQRO* **11**(1) (2018). ISSN: 2395–8847
 20. F.A. Jacques-García, D. Uribe-Mejía, G. Macías-Bobadilla, R. Chaparro-Sánchez, On modular inverse matrices, *XV Int. Eng. Congress*, ISSB:978-1-7281-5108-3/19 (2019)
 21. B. de Benito Crosetti, J.M.S. Ibáñez, La investigación basada en diseño en tecnología educativa. pp. 44–59. *Revista Interuniversitaria de Investigación en Tecnología Educativa (RIITE)*. ISSN: 2529–9638 (2016)
 22. R. McDermott, Experimental methodology in political science. *Political Analysis* 325–342 (2002)

A Comparative Study Between Two Numerical Methods for Symmetric Cryptography Uses and Applications

16

Alba Nidia Martínez-Martínez and Fausto A. Jacques García

Abstract

This document is focused in the comparison of two matrix numerical methods for symmetric cryptography, from a computational perspective in terms of memory, complexity and processing. The main task is to identify the most appropriate method along with Hill Cipher and form an improved cryptosystem. The methods are known as Gauss-Jacques and Gauss-Jordan with explicit modularization. Both of them could be used for the processing of secret keys in the approach to Shannon's Perfect Secrecy, which is of vital importance in terms of security and information protection. The experimental method is used to evaluate and analyze the behavior of each method in RAM consumption, computational complexity and processing, through their implementation in a functional language.

Keywords

Computational cost · Cryptosystem · Gauss-Jacques · Gauss-Jordan with explicit modularization · Hill Cipher · Key matrix · Modular inverse matrix · Numerical methods · Security information · Shannon's perfect secrecy · Symmetric cryptography

enced the direction and the means in how we communicate with each other and how we do anything in this evolving world [1].

With the rise in the use of computers and computer systems, information security is a priority in any digital environment. One of the most important activities is to keep the information that is transmitted and its storage safe [2].

There are security procedures with different objectives and forms. We have the physical ones that serve as a shield or those that seek to prevent a message from being intercepted by an attacker [3] and, cryptography is one of these procedures. Cryptography is a method that protects information using codes or symbols. It refers to the treatment of secret writing [4], and it has been used in several fields.

Cryptography has two classifications: classical and quantum cryptography. The first one is composed by different techniques, which are classified by the type of operation they use to transform the original text. For example, one algorithm or method is in which there is a substitution of letters or symbols. Another classification is based on how you process the original message. It means you can use a block or continuous processing. One more classifier is the number of keys you use to encrypt and decrypt, and they can be symmetric and asymmetric, respectively.

Symmetric or one-key cryptography uses a similar key during the encryption and decryption process [5]; that is, the private key at the sender is similar to the private key at the receiver (Fig. 16.1).

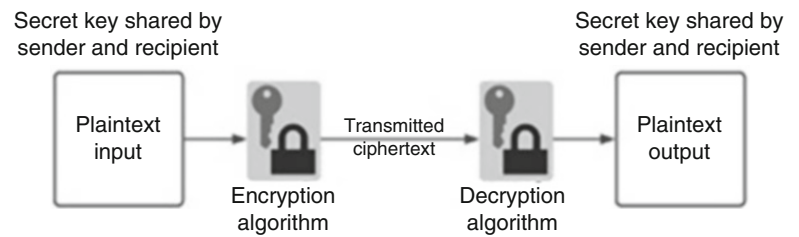
There are two factors in terms of security, which make a strong and secure cryptographic model. One of these factors is the secrecy of the key. If the key can be easily known by a cryptanalyst, the transmitted information is exposed and vulnerable. Due to the above, the key and its treatment is a critical point for the encryption and decryption processes in every method.

16.1 Introduction

For some years now, the use of digital media has intensified to establish oral and written communication, being it in a personal, educational, or even in the business environment. The obvious development of technology has drastically influ-

A. N. Martínez-Martínez (✉) · F. A. Jacques García
 Querétaro State University, Computer Science School,
 Querétaro, Mexico
 e-mail: amartinez361@alumnos.uaq.mx; jacques@uaq.edu.mx

Fig. 16.1 Symmetrical encryption process (source: own design)



A well known model of the use of numerical methods and cryptography is Hill Cipher. It was developed by Lester S. Hill in 1929.

In Hill Cipher, a number is linked to each letter or symbol of a language; it uses is made of basic arithmetic operations: addition, subtraction, multiplication, and division, but only with the integers of the corresponding module. The modulus is equal to the total of letters and symbols considered. It facilitates and reduces the operations that will be done to encrypt a message and to build a key. This technique uses a square matrix of numbers. The matrix is processed through a linear transformation to generate the key. The key and its command will be used to encrypt the message and send it. At this point, the key used must have the characteristic of being invertible [6].

Hill Cipher has some factors that can make it vulnerable. For example, a known-plaintext attack [7]. It is necessary to generate new contributions to treat the keys and improve the security.

Therefore, this study pretends the characterization of Gauss-Jacques and Gauss-Jordan with explicit modularization methods in order to identify the best use for each one. We also proposed to carry out a comparative analysis in three dimensions: RAM, computational complexity and general performance of the CPU.

However, it is not just a matter of adding security or generates a robust system against possible attacks but it is necessary to consider the resources that will be used. We have the modified Hill Cipher method, adding three stages which consist of 128-bit blocks and their 256-bit key. Although the result was a higher level of safety, the overall performance of the method did not improve. In other words, the method got slower [8].

When an algorithm is fast, or its processing requires fewer resources and has a low computational cost, we can call it an efficient method. The low computational cost will generally make an algorithm faster [9].

16.2 State of Art

16.2.1 Modular Inverse Matrices

A fundamental and basic part of matrix theory is found in the definition of the identity matrix and the inverse of a matrix.

In terms of a square $n \times n$ matrix, the identity matrix is a resultant matrix where the main diagonal is made up of the values ones at each position, and the rest of the elements in the matrix are zeros.

On the other hand, not all square $n \times n$ matrices will have their respective inverse matrix. When this happens, the original matrix is classified as singular; When a matrix is invertible, it is known as non-singular. An inverse matrix is unique [10].

So, since not all matrices have their inverse, it has been necessary to study which matrices will have it and how it could be found.

16.2.2 Gauss-Jacques Method

The Gauss-Jacques method, whose author is Doctor Fausto Abraham Jacques-García, uses the Gaussian elimination method on three linear axes. Furthermore, it does not use the determinant or the adjoint matrix to calculate the modular inverse of a matrix [11] instead, it follows this procedure:

1. Set the size of the key-matrix ($n \times n$).
2. Use RNG to generate the key-matrix.
3. Select modular value m as a prime number.
4. Write the identity matrix next to the candidate key-matrix found.
5. Find x , where $k_{ij}x$ is equivalent with 1 (mod m).
6. Apply the formula to the entire row, such as, $r_n x \bmod m = \text{new } r_n$. Where r_n is n -row.
7. Now, apply Gaussian elimination.
8. Repeat from step 5 to the next pivot.
9. The right-sided matrix is the modular inverse matrix.

The result of applying this method is a modular inverse matrix with no theoretical limit. The computational complexity expressed in Big O terms is $O(\phi^n)$. Equation (16.1) represents how it is computed:

$$K_m^{-1} = \left\{ \sum_{i=1}^s \sum_{j=1}^s [-k_{j(i+1)} + (k_{ij}e(k_{ii}, m) \bmod m)] \bmod m \right\} \quad (16.1)$$

16.2.3 Gauss-Jordan with Explicit Modularization

Another method to calculate inverse matrices is known as Gauss-Jordan and applied to cryptography, an explicit modularization is added at the end of the computation. As we know, the inverse matrix does not always exist. To evaluate if whether a matrix has an inverse matrix, the determinant must be different from zero. The steps to follow are:

1. Set the size of the key matrix.
2. Select the module (m).
3. Randomly generate the key matrix.
4. Write the key matrix on the left side and its identity matrix on the right side.
5. Perform basic operations, such as addition, subtraction, multiplication, and division. The identity matrix must be generated on the left side.
6. Once the identity matrix has been generated, the right side or the right matrix will be the inverse matrix.

Once the inverse matrix has been obtained, it is necessary to transform it into a matrix with integers of modulus m .

1. Find the inverse of the determinant y using the module m .
2. Multiply each numerator of the inverse matrix by y .
3. Apply the module m to each element of the obtained matrix.

After this procedure, the key is generated.

16.2.4 Uses and Applications

Among the main applications of modular arithmetic and inverse modular matrices in cryptography, one of them is the generation and treatment of the key matrix. It is necessary taking into account factors such as computational complexity, performance and resources consumed.

The Gauss-Jacques and Gauss-Jordan methods with explicit modularization have a series of well-defined steps to calculate the inverse modular matrix of the key matrix. They could be considered as algorithms and they can be used in different programming platforms so that organizations, researchers, students can use them for cybersecurity, cryptography, and encryption projects.

16.3 Purpose

This work has the general objective of a comparative analysis project between two numerical methods for cryptography: Gauss-Jacques and Gauss-Jordan with explicit modularization. The focus of the comparative study will be according to the following elements:

1. Characterization of the Gauss-Jacques method and Gauss-Jordan with explicit modularization method, that is, identify and explain its characteristics.
2. Compare three dimensions necessary for the analysis to offer useful codes in cryptography matters.

The dimensions considered for this analysis are:

1. Utilization of Random Access Memory, which is related to the performance of the device.
2. Computational complexity to identify the efficiency of each method to solve specific problems.
3. General performance in the Central Process Unit, at the processing level.

Therefore, each dimension allows the identification of different factors that are necessary to select and apply the most useful method when faced with a problem to be solved.

16.4 Methods

This comparative analysis proposes an experimental method. In the study of algorithms in computational sciences, one of the main tasks is to characterize them and make recommendations.

The suggested process is shown in Fig. 16.2.

16.5 Expected Results

The purposes that are being considered to identify the best application for both methods Gauss-Jacques and Gauss-Jordan with explicit modularization, are the following:

1. Characterization of the Gauss-Jacques method and Gauss-Jordan with explicit modularization method.

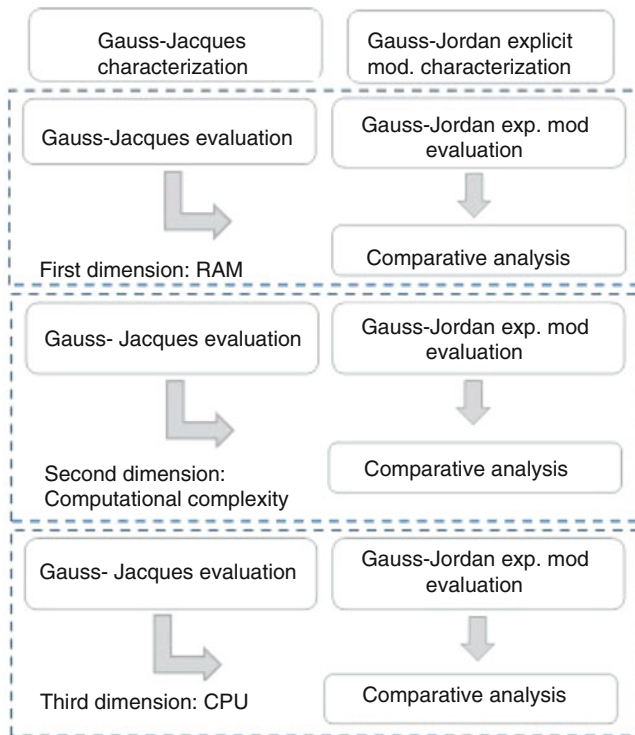


Fig. 16.2 Suggested process (source: own design)

2. Compare three dimensions necessary for the analysis to offer useful findings in cryptography matters.

Therefore, the results expected are:

1. Identify in which scenarios it is better to use Gauss-Jacques or Gauss-Jordan with explicit modularization.
2. Develop libraries for programming languages, in this case, GNU Octave.
3. Contribute to the computational economy in the treatment of the key matrix used to decrypt.

This document represents a proposal whose main objective is to contribute and expand the possibilities of cryptographic security.

References

1. K. Schwab, *The Fourth Industrial Revolution*. Currency (Klaus Schwab, New York, 2017)
2. H.R. Pawar, D.G. Harkut, Classical and quantum cryptography for image encryption & decryption, in *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)* (IEEE, New York, 2018), pp. 1–4
3. W. Stallings, *Cryptography and Network Security, 4/E* (Pearson Education India, New Delhi, 2006)
4. D.E.R. Denning, *Cryptography and Data Security*, vol. 112 (Addison-Wesley, Reading, 1982)
5. S. Vatschayan, R.A. Haidri, J.K. Verma, Design of hybrid cryptography system based on vigenère cipher and polybius cipher, in *2020 International Conference on Computational Performance Evaluation (ComPE)* (IEEE, New York, 2020), pp. 848–852
6. R. Ibañez, Criptografía con matrices, el cifrado de hill [Online]. Available: <https://culturacientifica.com/2017/01/11/criptografia-matrices-cifrado-hill/>
7. A. Meizar, F. Tambunan, E. Ginting et al., Optimizing the complexity of time in the process of multiplying matrices in the hill cipher algorithm using the strassen algorithm, in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, vol. 7 (IEEE, New York, 2019), pp. 1–4
8. A.A. Khalaf, M.S. Abd El-karim, H.F. Hamed, Proposed triple hill cipher algorithm for increasing the security level of encrypted binary data and its implementation using FPGA, in *2015 17th International Conference on Advanced Communication Technology (ICACT)* (IEEE, New York, 2015), pp. 454–459
9. H. Bercag, O. Kukrer, A. Hocanin, Recursive inverse adaptive filtering algorithm with low computational complexity on sparse system identification, in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (IEEE, New York, 2018), pp. 662–666
10. S.I. Grossman, *Álgebra Lineal* (McGraw Hill Educación, New York, 2008)
11. F.A. Jacques-García, D. Uribe-Mejía, G. Macías-Bobadilla, R. Chaparro-Sánchez, On modular inverse matrices, *XV International Engineering Congress* (2019). ISSN:978-1-7281-5108-3/19

Nir Drucker and Shay Gueron

Abstract

Rainbow is a Digital Signature Algorithm (DSA) that is based on multivariate polynomials. It is one of the Round-3 candidates of the NIST's Post-Quantum Cryptography Standardization project. Its computations rely heavily on $GF(2^8)$ arithmetic and the Rainbow submission optimizes the code by using AVX2 *shuffle* and *permute* instructions. In this paper, we show a new optimization that leverages: (a) AVX512 architecture; (b) the latest processor capabilities Galois Field New Instructions (GF-NI), available on Intel "Ice Lake" processor. We achieved a speedup of $2.43 \times / 3.13 \times / 0.64 \times$ for key generation/signing/verifying, respectively. We also propose a variation of Rainbow, with equivalent security, using a different representation of $GF(2^8)$. With this variant, we achieve a speedup of $2.44 \times / 4.7 \times / 2.1 \times$ for key generation/signing/verifying, respectively.

Keywords

Constant-time implementation · Galois field arithmetic · Multivariate polynomials · New processors' architectures · NIST PQC · Parallel computing · Performance · Post quantum cryptography · Rainbow · Signature schemes

17.1 Introduction

The potential threat to public key cryptography that large-scale quantum computers pose triggered the National Insti-

N. Drucker (✉) · S. Gueron
University of Haifa, Haifa, Israel
Amazon, Seattle, WA, USA

tute of Standards and Technology (NIST) to launch a standardization process for quantum-resistant crypto-algorithms [1]. This is currently a vibrant research topic. From the 69 Round-1 submission candidates only 4 Key Encapsulation Mechanisms (KEMs) and 3 DSAs made it to Round-3 of this project.¹ Rainbow [2] is one of these signature schemes and its security relies on the generic (NP-hard) Multivariate Quadratic (MQ) problem. It is a generalization of the Unbalanced Oil and Vinegar (UOV) scheme [3]. Rainbow enjoys a small signature size: 64/156/204 bytes for the Ia/IIIc/Vc variants, respectively. In addition, the signing and verifying operations are relatively quick. These features make it an appealing candidate. The submission includes several variants, from which we focus here on IIIc-Classic (the reasons are explained in Sect. 17.2.3).

The KeyGen/Sign/Verify computations of Rainbow rely on multiplications and inversions in $GF(2^8)$. A specific representation of $GF(2^8)$ as the $GF(2^{2^2})$ tower is used in the specification itself and we denote this representation by \mathbb{F}_{Tower} . The authors of Rainbow motivate this choice by the ease of a constant-time implementation of the code. However, we point out that any other field representation could also be used, with equivalent security.

In this paper, we explore the potential advantage that can be derived from a judicious use of new processor instructions in order to speedup Rainbow. Specifically, the GF-NI instructions [4] that are available on the latest $\times 86-64$ CPUs (microarchitecture codename "Ice Lake") [4]. The use of the GF-NI is demonstrated in [5] for some use-cases. Note that the GF-NI instructions operate over a specific representation of $GF(2^8)$ that was chosen for accelerating the symmetric encryption algorithm AES, we denote it by \mathbb{F}_{AES} . To leverage these instructions, we first need to calculate the conversion between \mathbb{F}_{Tower} and \mathbb{F}_{AES} . Furthermore, if one agrees to

¹Additional 5 KEMs and 3 DSAs were chosen by NIST as alternate candidates.

define Rainbow over \mathbb{F}_{AES} , conversion is no longer needed, and the implementation becomes faster.

The paper is organized as follows. Section 17.2 describes the new GF-NI instructions and the Rainbow signature scheme. We discuss the details of $\mathbb{F}_{\text{Tower}}$ and the conversion from/to \mathbb{F}_{AES} in Sect. 17.3. Section 17.4 discusses different implementation choices for Rainbow. Section 17.5 describes the experimental setup and Sect. 17.6 provides the performance results that we obtain. We conclude with Sect. 17.7.

17.2 Preliminaries

In this paper, we mark hexadecimal notation with a `0x` prefix, and place the LSB on the right-most position. For example, the byte string `0x11CC` is the binary string `0001000111001100`. Let X be a string of bits. We use $X[j : i]$, $j \geq i \leq 0$ to denote the sub-string of X that includes all the bits in the positions between i and j (included). We define $X[i : i] = X[i]$. For example, if $X = 000100011011$ we have $X[4 : 2] = 110$ and $X[7 : 7] = X[7] = 0$.

17.2.1 Galois Field Representation

Elements in Galois field can be represented in different ways, where all the representations are isomorphic. In particular, every representation in $GF(2^8)$ is a linear space of dimension 8 over $GF(2)$ and for every two representations $\mathbb{F}_X, \mathbb{F}_Y$ of $GF(2^8)$ there exists an isomorphism, a linear transformation that can be expressed by some (invertible) 8×8 binary matrix A :

$$\begin{aligned} \phi : \mathbb{F}_X &\longrightarrow \mathbb{F}_Y \\ x &\longmapsto A \cdot x \end{aligned}$$

We call A a conversion matrix. We denote by \mathbb{F}_{AES} the polynomial representation of $GF(2^8)$ with reduction polynomial $P_{\text{AES}} = x^8 + x^4 + x^3 + x + 1$. In this representation, multiplication in $GF(2^8)$ is done as standard polynomial multiplication reduced modulo P_{AES} . The bytes representation of P_{AES} is `0x11B`.

17.2.2 Vectorized GFNI

GF-NI includes the `VGF2P8MULB`, `VGF2P8AFFINEQB`, and `VGF2P8AFFINEINVQB` instructions. For short, we denote them by `MULB`, `AFFINEB`, and `AFFINEINVB`, respectively. Algorithm 1 describes `MULB`. It performs vectorized

multiplication (i.e., multiplying several elements in parallel) in \mathbb{F}_{AES} , of $KL = 16/32/64$ 8-bit elements that reside in two 128/256/512-bit registers (the registers are called `xmm`, `ymm`, `zmm`, respectively).

Algorithm 1 MULB instruction

Inputs: SRC1, SRC2 (wide registers)
Outputs: DST (a wide register)

```

1: procedure VGF2P8MULB(SRC1, SRC2)
2:   for j in 0 to (KL-1) do
3:     DEST.byte[j] ← GF2P8MULBYTE(SRC1.byte[j],
4:     SRC2.byte[j])
5:   end for
6: end procedure

7: procedure GF2P8MULBYTE(s1b, s2b) ▷ s1b,s2b (8 bits)
8:   T[15:0] = 0
9:   for i in 0 to 7 do
10:    if s2b[i] then
11:      T[15:0] = T[15:0] ⊕ (s1b ≪ i)
12:    end if
13:  end for
14:  for i in 14 downto 8 do
15:    if T[i] then
16:      T[15:0] = T[15:0] ⊕ (0x11b ≪ (i - 8))
17:    end if
18:  end for
19:  return T[7:0]
20: end procedure

```

We note that `MULB` can be used for different $GF(2^8)$ representations. This requires some conversions to/from these representations that can be performed with the `AFFINEB` (and `AFFINEINVB`) instruction described in Algorithm 2. Here, an affine transformation is $C \cdot x + b$ (or $Cx^{-1} + b$), for some 8×8 -bit matrix C that is “vectorized” $KL = 2/4/8$ times and for some 8-bit vectors x and b .

Algorithm 2 AFFINEB and AFFINEINVB instructions

Inputs: S1, S2 (wide registers) imm8 (8 bits)
Outputs: D (a wide register)

```

1: procedure VGF2P8AFFINE[INV]QB(S1, S2)
2:   for j in 0 to KL - 1 do
3:     for b in 0 to 7 do
4:       k = 64j, q = k + 8b
5:       D[q + 7 : q] = [Inv]AffB(S2[k + 63 : k], S1[q + 7 : q],
6:       imm8)
7:     end for
8:   end for
9:   return D[64KL - 1 : 0]
10: end procedure

11: procedure [Inv]AffB(s2, s1, imm8)
12:   for i = 0 to 7 do
13:     T[7-i] = parity(s2[8i+7 : 8i] & [inv](s1)) ⊕ imm8[i]
14:   end for
15:   return T[7:0]
16: end procedure

```

17.2.3 Rainbow

Rainbow is a multivariate-polynomial signature scheme defined over a finite field \mathbb{F} . It uses a system of m equations with n variables. Let us fix the number of layers u and to set $v_1, \dots, v_{u+1} \in \mathbb{Z}$ such that $0 < v_1 < \dots < v_{u+1} = n$. In addition set $V_i = \{1, \dots, v_i\}$ and $O_i = \{v_i + 1, \dots, v_{i+1}\}$, $i = 1, \dots, u$. Here, $m = n - v_1$, $|V_i| = v_i$ and set $o_i = |O_i|$. The Rainbow operations are as follows.

KeyGen The private key consists of two invertible affine maps $\mathcal{S} : \mathbb{F}^m \rightarrow \mathbb{F}^m$ and $\mathcal{T} : \mathbb{F}^n \rightarrow \mathbb{F}^n$, and a quadratic (invertible) central map $\mathcal{F} : \mathbb{F}^n \rightarrow \mathbb{F}^m$, consisting of m multivariate polynomials $f(v_1 + 1), \dots, f(n)$. The public key is the composed map $\mathcal{P} = \mathcal{S} \cdot \mathcal{F} \cdot \mathcal{T} : \mathbb{F}^n \rightarrow \mathbb{F}^m$ and therefore consists of m quadratic polynomials in the ring $\mathbb{F}[x_1, \dots, x_n]$.

Sign To sign a message m , compute its hash digest $h = H(m)$ with a hash function² $H : \{0, 1\}^* \rightarrow \mathbb{F}^m$. Compute $x = \mathcal{S}^{-1}(h) \in \mathbb{F}^m$ and its pre-image $y \in \mathbb{F}^n$ under the central map \mathcal{F} . Then, compute the signature $z = \mathcal{T}^{-1}(y) \in \mathbb{F}^n$.

Verify To verify a signature $z \in \mathbb{F}^n$ on a message m , calculate $h = H(m)$ and $h' = \mathcal{P}(z) \in \mathbb{F}^m$. Accept z if and only if $h' = h$.

Parameters Choice The Rainbow submission proposes three parameter sets in the form $(\mathbb{F}, v_1, o_1, o_2)$ as follows:

- Ia: $(GF(2^4), 32, 32, 32)$ with $m = 64$ equations and $n = 96$ variables. This is designed to meet NIST's security category Level-1/2.
- IIIc: $(GF(2^8), 68, 36, 36)$ with $m = 72$ equations and $n = 140$ variables. This is designed to meet NIST's security category Level-3/4.
- Vc: $(GF(2^8), 92, 48, 48)$ with $m = 96$ equations and $n = 188$ variables. This is designed to meet NIST's security category Level-5/6.

This paper focuses on the IIIc option, and the use of GF-NI to optimize its implementation. We note that it is possible to apply the proposed optimizations also to the Vc option.

Round-3 Rainbow Variants The Round-3 submission [6] adds two variants ("cyclic" and "compressed") to the Round-1 submission (called "standard"). As stated in [6], the KeyGen and Verify algorithms of cyclic-Rainbow are slower than the KeyGen and Verify of the standard-Rainbow. The compressed-Rainbow is similar to the cyclic-Rainbow and

the only difference is that it views the private key as a 512-bit seed. For this reason, it is enough to focus on standard-Rainbow.

17.3 Finite Field Representations for Rainbow

From the security viewpoint, the finite field representation used in [6] is immaterial. The specific choice of a tower field (\mathbb{F}_{Tower}) targets a constant-time implementation for the field multiplications. This representation views an element of $GF(2^8)$ as a degree-1 polynomial over $GF(2^4)$ as follows

- $GF(2^2) = GF(2)[e_1] = (e_1^2 + e_1 + 1)$
- $GF(2^4) = GF(2^2)[e_2] = (e_2^2 + e_2 + e_1)$
- $GF(2^8) = GF(2^4)[e_3] = (e_3^2 + e_3 + e_2e_1)$

Here, $GF(2^8)$ multiplication translates to $GF(2^4)$ operations, and these translate to $GF(2^2)$ operations. These can be easily executed in constant-time [7].

In particular, working in \mathbb{F}_{Tower} is convenient to program on a small device that can only perform $GF(2^2)$ multiplications (in constant-time). However, for typical modern server CPUs, different field representations are more appealing. Specifically, the use of \mathbb{F}_{AES} allows for leveraging the MULB instruction efficiently.

We outline several Rainbow flavors, based on different field representations.

- Working with \mathbb{F}_{Tower} : This requires conversion of input/outputs to/from \mathbb{F}_{AES} .
- Working with \mathbb{F}_{AES} : This does not require any conversion.
- Hybrid 1: The signing party stores the secret key in \mathbb{F}_{AES} and converts the signatures to \mathbb{F}_{Tower} . The verifying party stores the public key in \mathbb{F}_{Tower} .
- Hybrid 2: The signing party stores the secret key in \mathbb{F}_{Tower} and converts the signatures to \mathbb{F}_{AES} . The verifying party stores the public key in \mathbb{F}_{AES} .

The optimal choice depends on the computational power of the signing and verifying parties.

Conversion Across Field Representations As mentioned in Sect. 17.2, all the representations of $GF(2^8)$ are isomorphic. Therefore, it is possible to pass from one representation to another by means of multiplying by an 8×8 -bit matrix. The AFFINEB instruction is ideal for this purpose, and all that remains is to compute the conversion matrix and its inverse [8].

For our purposes we show how to compute the conversion matrix A from \mathbb{F}_{Tower} to \mathbb{F}_{AES} . We first choose a primitive element $\delta \in \mathbb{F}_{Tower}$ (e.g., $\delta = 0xbc$) such that δ is a root

²The concrete instantiation of Rainbow IIIc_Classic uses the SHA-384 algorithm as its hash function H .

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Fig. 17.1 The conversion matrix from \mathbb{F}_{Tower} to the \mathbb{F}_{AES}

$$A^{-1} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Fig. 17.2 The conversion matrix from \mathbb{F}_{AES} to \mathbb{F}_{Tower}

of P_{AES} (arithmetic in \mathbb{F}_{Tower}). Then, we compute the 8×8 binary matrix

$$A = [\delta^7, \delta^6, \delta^5, \delta^4, \delta^3, \delta^2, \delta^1, \delta^0]$$

with arithmetic in \mathbb{F}_{Tower} , where δ^0 is the multiplicative unit (i. e., $0x01$). Figures 17.1 and 17.2 show these matrices.

For using in `AFFINEB` the matrix A is represented by `0xf1f0a6869e3ab4ba` and the matrix A^{-1} by `0x03349c68700cdea0`.

17.4 Our Implementation

The official Round-3 implementation of Rainbow is found in [7]. It includes several variants one of which, is called “alternative”, uses AVX2 (technically C intrinsic) and is the fastest provided option. This “alternative” code performs fast GF multiplication of $a, b \in GF(2^8)$ by storing or calculating some multiplication tables using the same technique as described in [9, 10]. Subsequently, the tables are placed in 256-bit ymm registers and the multiplication is computed by 2 shuffle instructions (using `VPSHUF`), 2 AND instructions and one XOR instruction. By comparison, our implementation (available at [11]) simplifies the code because a multiplication involves only one `MULB` instruction with no tables at all. This also allows us to suggest further optimizations that are based on pipelining the code. Surprisingly, we found that

although modern compilers can unroll loops (automatically through a flag), hand written pipelining can still achieve faster results.

IIIc_Classic Dedicated Code The Rainbow implementation [7] supports multiple variants of Rainbow mentioned above with portable, SSE-based and AVX2 implementations. Our implementation is dedicated to the `IIIc_Classic` variant only. This facilitates dedicated optimizations for $\sigma_1 = \sigma_2 = 36$.

Inversion To perform inversion during `Sign`, we use `AFFINEINVB` as follows. We set $C = I$ and $b = 0$ so that `AFFINEINVB` computes $I \cdot x^{-1} + b = x^{-1}$. The hex representation of I is `0x0102040810204080`.

Using AVX512 The computations of Rainbow `IIIc_classic` operate on 72-byte rows, while AVX512 architecture has 512-bit zmm registers (64-bytes). Therefore, we use the AVX512 masking architecture that allows reading/writing only a part of a 512-bit zmm register. This saves some copies to/from temporary buffers and simplifies our code.

17.5 The Experimental Setup

The Platform For the experiments, we used a Dell XPS 13 7390 2-in-1 laptop. It has a 10th generation Intel® Core™ processor (microarchitecture codename “Ice Lake”[ICL]). The specifics are Intel® Core™ i7-1065G7 CPU 1.30GHz. This platform has 16 GB RAM, 48K L1d cache, 32K L1i cache, 512K L2 cache, and 8MiB L3 cache. For the experiments, we turned off the Intel® Turbo Boost Technology (in order to work with a fixed frequency and measure performance in cycles).

The Code We wrote the code mainly in C with some x86-64 assembly routines. The implementations use the GF-NI as well as other AVX512 instructions. We compiled the code with clang (version 9) in 64-bit mode, using the “-O3” optimization flag and ran it on a Linux OS (Ubuntu 18.04.2 LTS).

Remark 17.1 We note that GCC-8/9 also support the GF-NI instructions. However, during our study we identified a bug in GCC that causes incorrect results when using GF-NI. The bug is still present at the time of writing this paper. We reported it in [12] and the proper fix is underway.

Measurements Methodology The performance reported hereafter is measured in processor cycles (per single core),

Table 17.1 The performance of different implementations of Rainbow KeyGen/Sign/Verify. The numbers represent cycles count (in thousands), i. e., smaller is better. The code is profiled in the two measurement methodologies explained in Sect. 17.5

Implementation	KeyGen (10^6 cycles)		Sign (10^3 cycles)		Verify (10^3 cycles)	
	Method: Orig [7]	This work	Orig [7]	This work	Orig [7]	This work
Baseline	102	102	699	657	146	106
Baseline with CTR DRBG [13]	88.5 (1.16 \times)	88.5 (1.15 \times)	732 (0.95 \times)	675 (0.97 \times)	152 (0.96 \times)	106 (1.00 \times)
Impl1	42.1 (2.43 \times)	41.7 (2.44 \times)	264 (2.64 \times)	210 (3.13 \times)	226 (0.65 \times)	166 (0.64 \times)
Impl2	42.1 (2.43 \times)	41.7 (2.45 \times)	172 (4.05 \times)	142 (4.62 \times)	103 (1.41 \times)	56 (1.88 \times)
Impl3	42 (2.44 \times)	41.7 (2.45 \times)	168 (4.17 \times)	141(4.64\times)	106 (1.38 \times)	59 (1.81 \times)
Impl4	42 (2.43 \times)	41.8 (2.44 \times)	168 (4.17 \times)	143 (4.60 \times)	100 (1.47 \times)	50(2.13\times)

The bold numbers emphasize our best results. These are the numbers that we report in the paper abstract

where lower count is better. We obtain the results using two measurement methodologies.

- The methodology of [7]: Taking the average of 10 runs for the key generation, and the average of 500 runs for the Sign and Verify operations.
- Our methodology: Every measured function was isolated, run 25 times (warm-up), followed by 100 iterations that were clocked (using the RDTSC instruction) and averaged. To minimize the effect of background tasks running on the system, every experiment was repeated 10 times, and the minimum result was recorded.

The difference is in the minimization of background noise on the platform.

Code Packages Our baseline is the official “Alternative” code package that is submitted to the PQC project [7]. This implementation is written with AVX2 instructions. We compare it to our implementations of Rainbow:

- Impl1 - using GF-NI with elements in \mathbb{F}_{Tower} .
- Impl2 - using GF-NI with elements in \mathbb{F}_{AES} .
- Impl3 - using GF-NI with elements in \mathbb{F}_{AES} compiled with `-funroll-loops` clang flag.
- Impl4 - using GF-NI with elements in \mathbb{F}_{AES} compiled with `-funroll-loops` clang flag and manual pipelining optimization for Verify.

17.6 Results

Table 17.1 shows the performance results of our study. The first row shows the baseline, which is compared to our implementations in the subsequent rows. The heaviest operations in the key generation implementation are: (a) $GF(2^8)$ multiplications; (b) random number generation (noted already in [6]). To help isolating the performance contribution of GF-NI and AVX512 instructions we also replaced the DRBG of [7]

with our faster CTR DRBG implementation [13]. It is 1.16 \times faster.

For the Rainbow flavors that use \mathbb{F}_{AES} , we obtain a speedup factor of 4.64 \times for signing and 2.13 \times for verifying. Similar speedup is achieved for Rainbow flavors that use \mathbb{F}_{Tower} for signing. However, verifying is slowed down by a factor of (0.65 \times). This is due to the cost of converting the public key across from \mathbb{F}_{Tower} to \mathbb{F}_{AES} . This overhead can be eliminated by simply storing a copy of the public key in \mathbb{F}_{AES} (converting it only once).

The difference between Impl2 and Impl3 is very small. This indicates that adding the `-funroll-loops` compilation flag has a negligible effect (in this case). Note that manual pipelining achieves observable speedups with our measurement methodology (best versus average).

17.7 Conclusion

This paper shows how the new GF-NI instructions can be used for Rainbow `IIIC_classic`. We achieve speedups of 2.44 \times , 4.7 \times , and 2.1 \times for KeyGen/Sign/Verify, respectively, when the chosen field is \mathbb{F}_{AES} . This makes Rainbow a much more competitive candidate for the PQC standardization. Our results are measured on a laptop platform (the only platform with GF-NI that is currently available), and we expect to see even a stronger effect in future CPUs for server parts. We therefore recommend that the authors of Rainbow [6] consider a flavor of rainbow that operates in \mathbb{F}_{AES} as part of the modifications for Round-3.

The new optimized code of this paper is publicly available in [11].

Acknowledgments This research was supported by: The Israel Science Foundation (grant No. 1018/16); The Ministry of Science and Technology, Israel, and the Department of Science and Technology, Government of India; The BIU Center for Research in Applied Cryptography and Cyber Security, and the Center for Cyber Law and Policy at the University of Haifa, both in conjunction with the Israel National Cyber Bureau in the Prime Minister’s Office.

References

1. NIST, Post-Quantum Cryptography, <https://csrc.nist.gov/projects/post-quantum-cryptography> (2019). Last Accessed 20 Aug 2019
2. J. Ding, D. Schmidt, Rainbow, a new multivariable polynomial signature scheme, in *Applied Cryptography and Network Security*, vol. 3531, ed. by J. Ioannidis, A. Keromytis, M. Yung (Springer, Berlin, Heidelberg, 2005), pp. 164–175 [Online]. Available: https://doi.org/10.1007/11496137_1
3. A. Kipnis, J. Patarin, L. Goubin, Unbalanced oil and vinegar signature schemes, in *Advances in Cryptology — EUROCRYPT '99*, ed. by J. Stern (Springer, Berlin, Heidelberg, 1999), pp. 206–222 [Online]. Available: https://doi.org/10.1007/3-540-48910-X_15
4. Intel, Intel ®64 and IA-32 architectures software developer’s manual. System Programming Guide (2019)
5. N. Drucker, S. Gueron, V. Krasnov, The comeback of Reed Solomon Codes, in *2018 IEEE 25th Symposium on Computer Arithmetic (ARITH)*, June 2018, pp. 125–129. <https://doi.org/10.1109/ARITH.2018.8464690>
6. D. Jintai, C. Ming-Shing, P. Albrecht, S. Dieter, Y. Bo-Yin, Rainbow, March 2020 [Online]. Available: <https://csrc.nist.gov/projects/post-quantum-cryptography/round-3-submissions>
7. D. Jintai, C. Ming-Shing, P. Albrecht, S. Dieter, Y. Bo-Yin, Rainbow (IIIc_Classic) Alternative code package, March 2019. <https://csrc.nist.gov/CSRC/media/Projects/Post-Quantum-Cryptography/documents/round-2/submissions/Rainbow-Round2.zip>
8. S. Gueron, International NI: using AES-NI for fast and constant time Chinese SM4 and other ciphers, IACR ePrint, 2020
9. J.S. Plank, K.M. Greenan, E.L. Miller, Screaming fast galois field arithmetic using intel SIMD instructions, in *11th USENIX Conference on File and Storage Technologies (FAST 13)* (USENIX Association, San Jose, CA, 2013), pp. 298–306 [Online]. Available: https://www.usenix.org/conference/fast13/technical-sessions/presentation/plank_james_simd
10. S. Gueron, M. Kounavis, Efficient implementation of the galois counter mode using a carry-less multiplier and a fast reduction algorithm. *Inf. Process. Lett.* **110**(14), 549 – 553 (2010) [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S002001901000092X>
11. N. Drucker, S. Gueron, GFNI based Rainbow (IIIc_Classic), January 2020. <https://github.com/aws-samples/rainbow-with-gfni>
12. N. Drucker, GCC-8 considers the _mm_gf2p8affine_epi64_epi8 intrinsic to be symmetric, December 2019. <https://www.mail-archive.com/gcc-bugs@gcc.gnu.org/msg632510.html>
13. N. Drucker, S. Gueron, Fast CTR DRBG for x86 platforms, March 2019. <https://github.com/aws-samples/ctr-drbg-with-vector-aes-ni>

Extending a Hybrid Security Risk Assessment Model with CWSS

Robert Banks, Jim Jones, Noha Hazzazi, Pete Garcia,
and Russell Zimmermann

Abstract

Cybersecurity risk management is the foundation of business and organizational decisions involving digital technology. Various models have been proposed and are in use, but these apply to current technologies and use cases, and none are sufficient to evaluate new technologies. This paper builds upon prior work using CVSS to quantify potential security threats for which information is limited. That prior work merges CVSS data with MITRE's Common Attack Pattern Enumeration and Classification (CAPEC™) tools to inform a new technology risk scoring system in a Bayesian Belief Network (BBN). This work extends this risk model to incorporate CWSS data to better reflect the environments' weaknesses that may apply to new technologies. This approach enables a more accurate and trustworthy way of quantitatively estimating risk as a function of the Base Finding Subscore and Attack Surface

Subscore for weaknesses most relevant to businesses, missions, and deployed technologies.

Keywords

Bayesian belief network (BBN) · Common attack pattern enumeration and classification (CAPEC™) · Common weakness risk analysis framework (CWRAF) · Common weakness scoring system (CWSS) · Cyber survivability endorsement (CSE) · Generation of security · National Vulnerability Database (NVD) · Risk estimation model · Risk management and sensitivity analysis

R. Banks (✉)

Volgenau School of Engineering, George Mason University, Fairfax, VA, USA
e-mail: rbanks3@gmu.edu

J. Jones

Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA
e-mail: jjonesu@gmu.edu

N. Hazzazi

Department of Electrical and Computer Science, Howard University, Washington, DC, USA
e-mail: Noha.Hazzazi@howard.edu

P. Garcia

PricewaterhouseCoopers, LLP, Miami, FL, USA
e-mail: Pete.Garcia@pwc.com

R. Zimmermann

SAIC, Chantilly, VA, USA
e-mail: Russell.F.Zimmermann@saic.com

18.1 Introduction

Risk management is a useful tool for controlling risk, although it has limitations when trying to produce quantitatively accurate, reliable estimates of exploitability and impact. This paper builds upon a CVSS-based model for quantifying potential security threats, where data is often limited and rarely reusable because it involves confidential information. Risk estimation models are more effective and broadly useable if they are based on publicly available data sources. Our work uses MITRE's public Common Attack Pattern Enumeration and Classification tools [1] and the Department of Homeland Security's public National Vulnerability Database [2] risk measurements of technology vulnerabilities to populate a Bayesian Belief Network (BBN). We extend this prior work by incorporating the CAPEC CWSS data and tools to enable accurate and trustworthy quantitative estimates of the Base Finding and Attack Surface Subscore of new technology's weaknesses [1]. These two elements contribute to an overall risk model calculation.

The contributions of this methodology are three-fold:

1. A quantitative security risk level estimation for the next generation of security technologies using the CWSS for weakness most relevant to businesses, missions, and deployed technologies.
2. Use of the publicly available NVD data to provide probability distributions for the CWSS analysis. The CWSS leverages the NVD data as input distributions and constructs a model for analysis within a BBN structure.
3. A BBN model allows other sources to combine estimates with the CWSS metrics and various levels of abstraction for input information.

18.2 Problem Context

Several generations of cyber defenses represent cybersecurity research, but as technology weakness and dependence continue to grow, the number of incidents and their effects also grow. Partha Pal presents the evolution of cyber's first three Generations of Security Research [3]. He demonstrates that constantly changing but routine engineering designs result in increased volume, velocity, and variety of security incidents because these defensive capabilities are still operating in a vertical/silo mindset. Pal's first-generation prevented intrusion through protection, e.g., via robust cryptographic algorithms. The second-generation intended to detect previous preventions' failures to contain their effects, e.g., through network intrusion detection systems and anti-virus tools. The current practice is the third-generation of security, which focuses on survivability, developing systems that can tolerate and recover. Simultaneously, defenses improve over time from the impact of cyberattacks and hence can regain lost capabilities. These approaches provide limited individual benefits in today's "daily breach" environment, which has led to the fourth-generation cybersecurity research that hopes to change an attacker's cost/benefit calculus [4]. Despite these advances, cybersecurity research still needs to focus on weaknesses most relevant to businesses, missions, and technologies, especially as these businesses, missions, and technologies change.

18.3 Related Works

Relevant prior risk management work focuses on different aspects that reduce the CVSS vulnerability (FIRST, 2019) for a given CWSS weakness, use Commercial-Off-The-Shelf (COTS) software, employ fuzzy risk analysis, improve control measures, or improve system survivability.

The CVSS Risk Estimation Model is a security trade-off decision-support engine for balancing security with a cost. As discussed, the cost-benefit perspective of a security reflects financial and project factors such as

budget and time-to-market [5]. However, accurate risk level measurement remains a challenge. The availability security attributes outlined in the [6] estimation model form the basis for system service levels. This model relied on experts to define the service levels and included no evidence aggregating prior vulnerability distributions for future weaknesses.

Chen et al. discussed using CVSS for COTS software systems to measure security investment benefits [7, 8]. However, CWSS does not take the context values into account and could be misleading. The CWSS granularity supports the stakeholders' different perceptions to the extent that weakness might affect them. Chen's work for productivity and reputation are subjective and equally hard to estimate, as are their environmental metric group attributes.

Dondo's approach used fuzzy risk analysis for vulnerability prioritization to derive the risk level or risks to a system where the asset value is assumed to be known [9]. Asset value is specific to the context and stakeholder, which is not necessarily easy to evaluate as are the temporal and environmental metric group attributes.

NIST Special Publication 800-53 presents a proactive and systemic approach to developing comprehensive safeguarding measures [10]. The calculated risks remain a vital deficiency of these and other security risk assessments. The CWSS also lacks prior knowledge for determining an acceptable risk based on budget, time, and resource constraints.

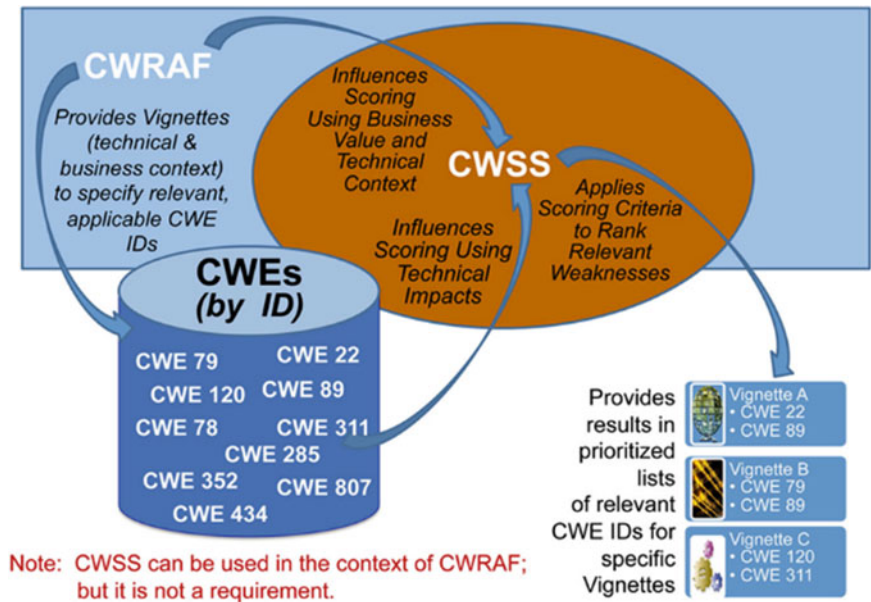
The Cyber Survivability Endorsement (CSE) is the critical foundation for ensuring Cyber Survivability Attributes (CSAs) are considered part of the operational risk trade-space [11]. CSE leverages the NIST 800-53 cybersecurity technical controls and does not define any new cybersecurity requirements. The CSE provides a holistic approach to determine a system's Cyber Survivability Risk Category (CSRC), and then to assess and manage its Cyber Survivability Risk Posture (CSRP) throughout its lifecycle. CSE determines acceptable risks based on budget, time, and resource constraints limited to the most recent activity and would benefit from prior vulnerability distributions applied to future weaknesses.

18.4 Methodology

This paper builds upon the Hybrid Security Risk Assessment that combines disparate information sources. This prior work applied the NVD for the CVSS probability distributions to implement a BBN model. This work extends the NVD and CAPEC CWSS tools in a different Bayesian model to assess next-generation software security [12].

The CAPEC CWSS (see overview in Fig. 18.1) provides information to enhance security throughout a software development lifecycle. The publicly available catalog in CWSS enables users to understand how adversaries exploit appli-

Fig. 18.1 CWSS overview



ation weaknesses and other cyber-enabled capabilities [13]. The CAPEC mechanism, CWSS, scores these weaknesses in a consistent, flexible, open manner enabling an organization to reflect its business domain(s) [13].

Because CWSS normalizes the approach for characterizing weaknesses, CWSS users can invoke base findings, attack surface, and environmental metrics to apply contextual information that reflects the software risk more accurately in the unique business context within which it will function. This unique business capability enables stakeholders to make more informed decisions when mitigating risks posed by weaknesses.

18.5 Implementation

The Common Weakness Enumeration compares the CVSSv2 factors with CWSS Factors in Table 18.1 [13]. The comparison combines the multiple characteristics split into distinct factors. Nevertheless, there are several differences between these scoring systems. The CVSS assumes a vulnerability has been discovered and verified, and CWSS does not account for incomplete information. CVSSv2 has a more significant bias to the physical system, where the CWSS has less, and CVSSv3 removes this bias. Lastly, the CVSS and CWSS scores are not necessarily comparable due to differences in scales.

The Hybrid Security CWSS Risk Assessment Model leverages Table 18.1 to combine verified vulnerability data and account for incomplete information. This approach bypasses differences in scoring systems with a ratio of the calculated value of prior knowledge required in the BBN models.

Table 18.1 CVSS – CWSS crosswalk

Vulnerability	Weaknesses
CVSSv2	CWSSv1.0.1
Access Vector (AC)	Attack Vector (AC)
Access Complexity (AC), Target Distribution (TD)	Deployment Scope (SC), Acquired Privilege Laver (AL)
Attack Complexity (AC)	Required Privilege Level (RL), Acquired Privilege (AP), Required Privilege (RP)
User Interaction (UI)	Level of Interaction (IN)
Confidentiality Requirements (E_CR)	Technical Impact (TI)
Integrity Requirements (E_IR)	Technical Impact (TI)
Availability Requirements (E_AR)	Technical Impact (TI)
N/A	Authentication Strength (AS)
Remediation Level (RL)	Internal Control Effectiveness (IC)
Attack Complexity (AC)	External Control Effectiveness (EC)
Report Confidence (RC)	Finding Confidence (FC)
Confidentiality Impact (C)	Technical Impact (TI)
Integrity Impact (I)	Technical Impact (TI)
Availability Impact (A)	Technical Impact (TI)

18.6 BBN Topology

The BBN topology provides a seamless aggregation of CWSS weaknesses applied to new technologies’ security in a consistent, flexible, and open manner and reflecting an organizational context and business domain [14]. Additionally, the BBN provides a mathematically accurate way of assessing the effects of different events (or nodes, in this context) on each other. These assessments made in either direction reflect the most likely effects given the values

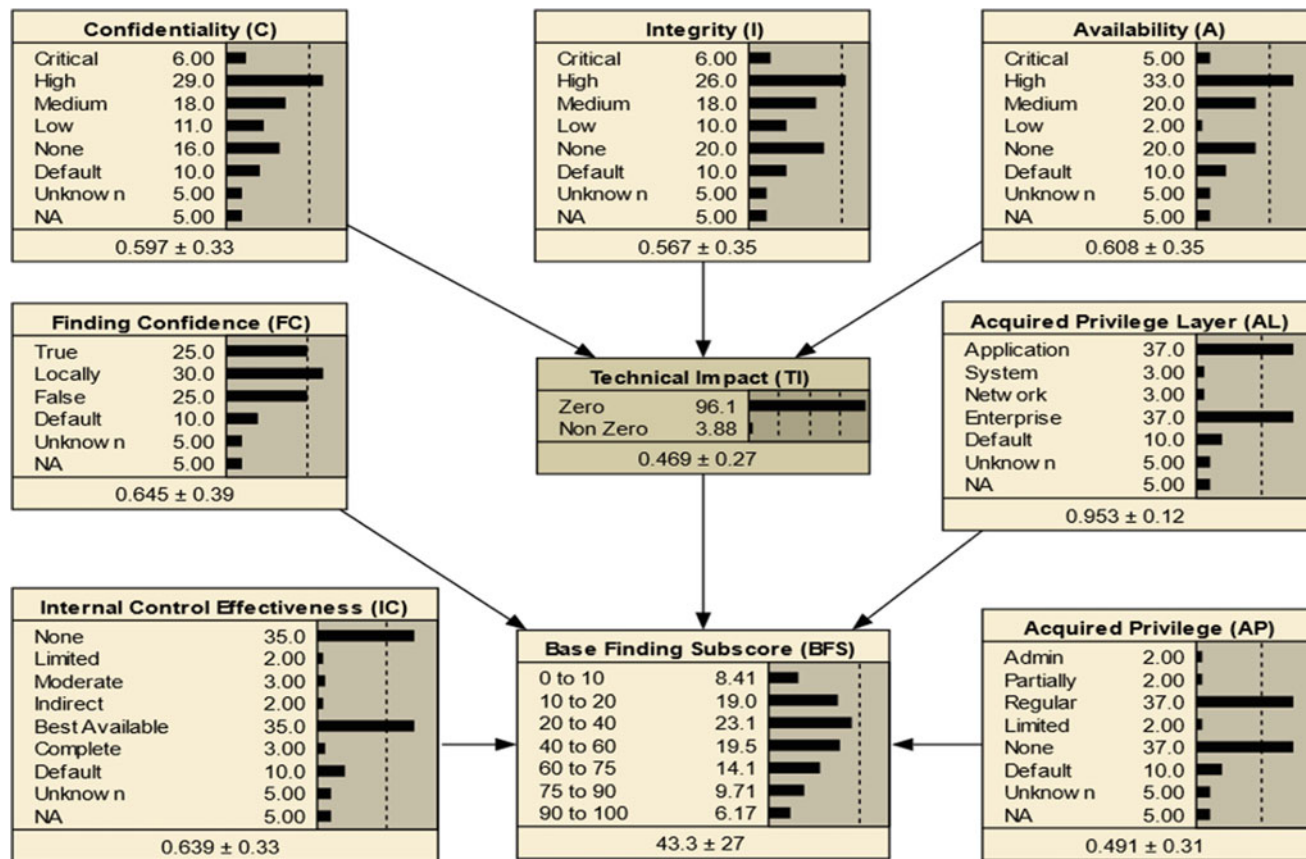


Fig. 18.2 CWSS BBN base finding subscore

of specific causes and determine the most likely causes of observed events.

The BBN models allow multiple layers of abstraction for the input information in order to derive exploitation, impact, and risk level estimates. The nodes and their states are derived from the MITRE CWSS, and their prior probability distributions are derived from the NVD data. The CWSS BBN models in Figs. 18.2 and 18.3 capture part of a new technology risk score computation by applying vulnerability information of existing technology to a CVSS baseline.

The BBN is an interactive model to select states that uses the CWSS Base Finding Subscore in Eqs. (18.1) and (18.2) and Attack Surface Subscore in Eq. (18.3) to produce child node results. The Base Finding or Attack Surface node belief bars show each state conditional probabilities.

CWSS standardizes the approach for characterizing weaknesses. Users of CWSS can invoke attack surface and environmental metrics to apply contextual information that more accurately reflects the risk to the software capability, given the unique business context in which it will function [13]. This unique business capability provides a Base Finding metric group that captures the inherent risk of the weakness

in the finding accuracy. An Attack Surface metric group contains barriers that an attacker must overcome to exploit the weakness [13]. The Environmental metric group contains characteristics of the weakness that are specific to an environment or operational context. CWSS is used in cases where there is little information at first, but the quality of information can improve over time.

Base Finding Subscore

$$TI(C, I, A) = (C + I + A) / 3 \quad (18.1)$$

$$BFS(TI, AP, AL, FC, IC) = TI == \text{Unchanged?} \\ ((10 * TI + 5 * (AP + AL) + 5 * FC) * 0 * IC) * 4.0 : \\ ((10 * TI + 5 * (AP + AL) + 5 * FC) * IC) * 4.0 \quad (18.2)$$

Attack Surface Score

$$ASS(RP, RL, AV, SC, IN, AS) \\ = ((20 * (RP + RL + AV) \\ + (20 * SC) + (15 * IN) + (5 * AS))) \quad (18.3)$$

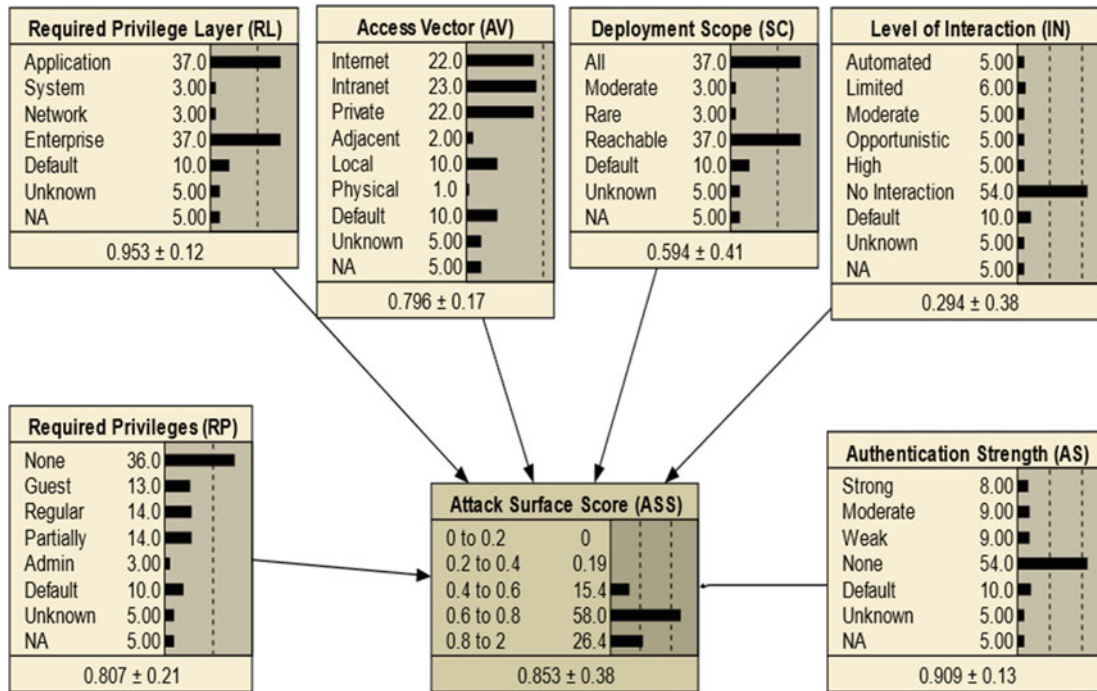


Fig. 18.3 CWSS BBN attack surface subscore

Table 18.2 Base finding subscore sensitivity of findings

Node	Variance reduction	Percent	Mutual info	Percent	Variance of beliefs
Acquired Privilege (AP)	0	0	0.00000	0	0.0000000
Availability (A)	0.1356	0.0172	0.00063	0.023	0.0000108
Confidentiality (C)	0.1586	0.0201	0.00073	0.0268	0.0000127
Integrity (I)	0.123	0.0232	0.00083	0.0307	0.0000146
Acquired Privilege Layer (AL)	0.8024	0.102	0.00224	0.0826	0.0000759
Technical Impact (TI)	2.081	0.264	0.00830	0.306	0.002134
Finding Confidence (FC)	7.673	0.975	0.01786	0.658	0.0004992
Internal Control Effective (IC)	165.3	21	0.36168	13.3	0.0247780
Base Finding Subscore (BFS)	787	100	2.71267	100	0.7058946

18.7 Sensitivity of Findings

A sensitivity analysis reveals how much a single finding could influence the target node’s beliefs at each of the other nodes in the net [15]. The sensitivity analysis reports the minimum and maximum beliefs for each state as 0 and 1, respectively, and the maximum reductions in variance and entropy will be 100%. The expected reduction in the query variable variance has an expected real value, as shown in Tables 18.2 and 18.3 for each variable, where a higher value is better. The mutual information is the expected reduction in the query variable’s entropy due to a finding at one or more of the varying variables and a higher value is better. A variance of beliefs is the expected change (squared) of the query variable’s beliefs, taken over all its states, due to

a finding at one or more varying variables. A lower value is the least disruptive.

When the sensitivities are calculated, all findings currently entered in the network will significantly affect the sensitivities. In the summary tables, we have selected the lowest state values to display their range in Tables 18.4 and 18.5 and the highest state values to display their range in Tables 18.6 and 18.7.

18.8 Discussion

The NVD files provide probability distributions for past vulnerabilities that can apply to next generation security technologies. The NVD data contains independent subgroups aggregated separately, which provide the distributed inputs for

Table 18.3 Attack surface subscore sensitivity of findings

Node	Variance reduction	Percent	Mutual info	Percent	Variance of beliefs
Authentication Strength (AS)	0.0009438	0.506	0.00554	0.524	0.0017971
Required Privilege Layer (RL)	0.005969	3.2	0.04082	3.86	0.009312
Access Vector (AV)	0.006361	3.29	0.04320	4.12	0.0108894
Required Privileges (RP)	0.008302	4.3	0.06427	6.13	0.0097766
Deployment Scope (SC)	0.08443	44.7	0.61592	53.8	0.1445353
Attack Surface Score (ASS)	0.1932	100	1.04761	100	0.2530914

Table 18.4 Base finding subscore lowest state probabilities nodes selected

Node	State	New finding	All findings
Confidentiality (C)	Low	11%	
Integrity (I)	Low	10%	1.1%
Availability (A)	Low	2.0%	0.022%
Finding Confidence (FC)	False	25.0%	0.0055%
Acquired Privilege Layer (AL)	Network	3.0%	0.00165%
Internal Control Effective (IC)	Best available	35%	5.775e-05%
Acquired Privilege (AP)	None	15.2353%	8.79838e-06%

Note: With a Base Finding Subscore (BFS) 10.2 Mean, +/-0.0 Std Dev., 10.2 Median, 0.0 IQR

Table 18.5 Attack surface score lowest state probabilities nodes selected

Node	State	New finding	All findings
Required Privilege Layer (RL)	Network	3.0%	
Access Vector (AV)	Physical	1.0%	0.03%
Deployment Scope(SC)	Reachable	37.0%	0.0111%
Level of Interaction (IN)	High	5.0%	0.000555%
Required Privileges (RP)	Admin	3.0%	0.665e-05%
Authentication Strength (AS)	Strong	8.0%	1.32e-06%

Note: With an Attack Surface Score (ASS) 0.27 +/- 0.0 Std Dev., 0.27 Median, 0.0 IQR

the CWSS factors within the BBN models. Previous work did not include past vulnerability distributions that could apply as a ratio for the Base Finding and Attack Surface weaknesses. Our methodology takes advantage of the publicly available data and presents a security risk level estimation for the next generation of security technologies that use the CWSS for weaknesses most relevant to businesses, missions, and deployed technologies.

The NVD publicly available data influences the CWSS analysis, and the BBN sensitivity analysis shows how much findings at other nodes influence an output node. Tables 18.2 and 18.3 show the difference in an expected reduction in variance and entropy of the expected output value due to a finding. Where Internal Control Effective and Deployment Scope are the most significant nodes, simultaneously the

Table 18.6 Base finding subscore highest state probabilities nodes selected

Node	State	New finding	All findings
Confidentiality (C)	Low	29%	
Integrity (I)	Low	26%	7.54%
Availability (A)	Low	5.0%	0.377%
Finding Confidence (FC)	Locally	30%	0.00131%
Acquired Privilege Layer (AL)	System	3.0%	0.003393%
Internal Control Effective (IC)	Limited	2.0%	6.786e-05%
Acquired Privilege (AP)	Partially	1.29412%	8.78188e-07%

Note: Technical Impact (TI) Zero State Selected with Probability of new finding = 100%, of all findings = 8.78188e-07% with a Base Finding Subscore (BFS) 80.4 Mean, +/-0.0 Std Dev., 80.4 Median, 0.0 IQR

Table 18.7 Attack surface score highest state probabilities nodes selected

Node	State	New finding	All findings
Required Privilege Layer (RL)	System	3.0%	
Access Vector (AV)	Intranet	22%	0.66%
Deployment Scope (SC)	Moderate	3.0%	0.0198%
Level of Interaction (IN)	Limited	6.0%	0.001188%
Required Privileges (RP)	Guest	13%	0.00015444%
Authentication Strength (AS)	Weak	9.0%	1.38996e-05%

Note: With an Attack Surface Score (ASS) 0.88 +/- 0.0 Std. Dev., 0.88 Median, 0.0 IQR

variance of beliefs provided the expected change (squared) over all its states, due to a finding. This analysis quantifies the influence at one node from findings at other nodes.

18.9 Conclusion

Our method presents a quantitative security risk level estimation for the next generation of security technologies that use the CWSS for weaknesses most relevant to businesses, missions, and deployed technologies. The results in Table 18.2 shows that the variable Internal Control Effective and Finding Confidence (FC) has the most influence on the variance reduction with the highest variance reduction, highest mutual information and lowest variance of belief. Further-

more, Availability, Confidentiality, and Integrity are less significant to a varying degree, and are the least important with respect to the evaluation criteria. The results in Table 18.3 show that variable Deployment Scope followed by Required Privileges have the most influence in the variance reduction with the highest mutual information, highest percent differences, and lowest variance of belief. Required Privilege Layer and Authentication Strength are the least important with respect to the evaluation criteria. The BBN sensitivity analysis shows the importance of the Hybrid Security CWSS Risk Assessment Model factors based on how much findings at other nodes influence a node.

Our methodology takes advantage of the publicly available NVD data to provide probability distributions for the CWSS analysis. NVD CVE incidents provide the prior knowledge necessary for the BBN.

The BBN topology allows other sources to combine with estimates of the CVSS information and various abstractions of input information [16]. The sensitivity analysis determines the most informative nodes in forming the beliefs of the most probable output nodes. That changes as findings arrive, so the model recomputes at each input.


In future work, the CVSS distributions of known vulnerabilities can also provide vignettes with prior knowledge in CWRAF-based estimates for BBN's for the Critical Infrastructure. Our method would facilitate a collaborative, community-based effort addressing stakeholders' needs, primarily when used in conjunction with the Common Weakness Risk Analysis Framework (CWRAF).

References

1. CAPEC – About CAPEC. (2019). <https://capec.mitre.org/about/index.html> (February 6, 2020)
2. NVD – Data Feeds, 2020. <https://nvd.nist.gov/vuln/data-feeds> (February 6, 2020)
3. P. Pal, R. Schantz, M. Atighetchi, J. Loyall, F. Webber, What next in intrusion tolerance. BBN Technologies, Cambridge. (2009). <http://wraits09.di.fc.ul.pt/wraits09paperParthaPal.pdf>, 8 citations
4. S. Jajodia, A. K. Ghosh, V. Swarup, C. Wang, X. S. Wang (eds.), *Moving Target Defense: Creating Asymmetric Uncertainty for Cyber Threats, 2011 edition* (Springer, 2013). <https://www.springer.com/gp/book/9781461409762>
5. S.H. Houmb, V.N.L. Franqueira. Estimating ToE risk level using CVSS. *Proceedings of the Fourth International Conference on Availability, Reliability and Security (ARES 2009 – The International Dependability Conference)*, pp. 718–725. (2009). <https://doi.org/10.1109/ARES.2009.151>, 52 citations
6. S.H. Houmb, V.N.L. Franqueira, E.A. Engum, Quantifying security risk level from CVSS estimates of frequency and impact. *J. Syst. Softw.* **83**(9), 1622–1634 (2010) <http://www.sciencedirect.com/science/article/pii/S0164121209002155> (October 20, 2019)
7. Y. Chen, Stakeholder value driven threat modeling for off the shelf based systems. *29th International Conference on Software Engineering (ICSE'07 Companion)*, pp. 91–92. (2007). <https://doi.org/10.1109/ICSECOMPANION.2007.69>, 21 citations
8. Y. Chen, B. Boehm, L. Sheppard. Measuring security investment benefit for off the shelf software systems – a stakeholder value driven approach. *Online Proceedings of Sixth Workshop on the Economics of Information Security (WEIS 2007)*. (2007). <http://weis07.infosecnet.net/papers/46.pdf>
9. M.G. Dondo, A vulnerability prioritization system using a fuzzy risk analysis approach. *Proceedings of The Ifip Tc 11 23rd International Information Security Conference*, pp. 525–540. https://doi.org/10.1007/978-0-387-09699-5_34, 22 citations (2008)
10. T. Allen, NIST Cybersecurity for IoT Program – Publications. NIST. (2019). <https://www.nist.gov/itl/applied-cybersecurity/nist-cybersecurity-iot-program/publications> (August 23, 2020)
11. J. Petty, Cybersecurity Test and Evaluation Guidebook 2.0. (2018). https://daytonaero.com/wp-content/uploads/DOD_Cybersecurity-Test-Evaluation-Guidebook-ver-2.0_25-APR-2018.pdf
12. O. Pourret, P. Naïm, B. Marcot, *Bayesian Networks: A Practical Guide to Applications* (Wiley, 2008) 443 citations. <https://vbn.aau.dk/en/publications/an-introduction-to-bayesian-networks>
13. CWE – Common Weakness Scoring System (CWSS). (2014). https://cwe.mitre.org/cwss/cwss_v1.0.1.html (February 6, 2020)
14. Introduction to Bayesian Networks—Finn V. Jensen—Google Books. 1997, 4621 citations
15. Norsys, Tutorial on Bayesian Networks with Netica – Basics. (2020). https://www.norsys.com/tutorials/netica/nt_toc_B.htm
16. CVSS v3.1 Specification Document, FIRST — Forum of Incident Response and Security Teams. (2019). <https://www.first.org/cvss/specification-document>

Part IV
E-Health

Identifying and Prioritizing Applications of Internet of Things in the Supply Chain of Distribution and Sale of Health Care Products in Iran

Niloofar AminiKalibar and Fatemeh Saghafi 

Abstract

Applying Internet of Things (IoT) leads to improvements in the quality of life. Supply chains are considered as the world's most fundamental parts, for increasing productivity of which, countless efforts are applied. Although rare studies are conducted on the IoT's application, no study was found on the distribution and sale part of the supply chain. Solving challenges in this area is lucrative for service provider companies since products of this industry have a considerable share in the family shopping basket. Firstly, properties and advantages of the IoT and its maturity level are studied. Since technology implementation depends on the context, the challenges of health care products' supply chain in the distribution and sale sectors along with corresponding solutions to overcome such challenges are addressed through interviews with 23 experts in two groups of IoT and Health Care Products Industry. Interviews are analyzed by the content analysis method. As implementing such technologies are not practicable unless the technology is matured in the country, each application is prioritized according to its corresponding technology maturity. The result is applicable for the Health Care Industry and investors aiming to develop their new technology-based industry.

Keywords

Internet of things · E-commerce · Sale · Distribution · Supply chain · Health care products · Maturity · IoT application · Technology implementation · Content analysis

N. AminiKalibar (✉) · F. Saghafi
 Faculty of Management, University of Tehran, Tehran, Iran
 e-mail: amininiloofar@ut.ac.ir; fsaghafi@ut.ac.ir

19.1 Introduction

Today, sharp competition in industries proceeds in global markets. Supply chains, prominent roles in this competitive environment, are driving the business flow. Applying the technology of Internet of Things has major role in establishing productivity and comforting life for human beings. On the other hand, Internet of Things is a kind of green energy; the implementation of which avoids any pollution to the environment that can happen while running supply chains. In fact, Carbon production will be cratered in the environment [1].

Supply chain management (SCM) means providing correct products in the correct volume, time, place, price, and condition to a correct customer [2]. There were challenges in traditional supply chains such as inaccurate volumes, delays, shortage, and overstock rising from uncertainty, human errors, and complexity. Smart supply chains are capable of overcoming described challenges. A smart supply chain is an interconnected system covering separated applications to regional and integrated ones [3]. Improvement in the supply chain is possible by the means of information technology as it has the potential to integrate processes.

The Internet of Things is the logical extension of the internet to the physical world; it is going to establish major business model development. Furthermore, it might even be more disruptive than the internet in the 90s as it reaches beyond computers to basically any item in the world [4].

The flow of information and Big Data increases the value of IoT exponentially. Starting with an active but non-connected device, a completely cross-ecosystem and integrated automation is at the end of the maturity curve. The deployment strategy of Internet of Things depends on the environment and the country in which the technology is applied [1]. Therefore, implementing the technology of

Internet of Things is not practicable unless right platform is provided for the technology.

Health care products have substantial share in the family shopping basket since attention to hygiene and beauty comprises paramount importance in peoples' life. It is the reason why this industry is formed and has achieved excellent growth. The purpose of this research is identifying and prioritizing the utilization of IoT in health care products' supply chain in Iran. To conduct it, in the second section (literature), Internet of things and its application in a supply chain is introduced. Afterwards, maturity layers of IoT are demonstrated. Following that, articles on the application of IoT in a supply chain is reviewed and the research gap is defined. In the third section research methodology is indicated. Findings and conclusion are discussed in the fourth and fifth section.

19.2 Literature

19.2.1 Internet of Things and Its' Application in Supply Chain Management

Supply chain management (SCM) is the management process of supply chain activities to meet customer requirements most efficiently and achieve a sustainable competitive benefit.

In traditional supply chain management systems had serious challenges such as overstocking, delivery delays, shortage, inventory lost, and uncontrolled problems all of which referred to factors such as complexity, invisibility, and uncertainty. Information technology has jumped on to defeat the obstacles of supply chain management systems. IoT has the correct potential to fill the barriers mentioned.

Internet of things can be defined in several ways; its system includes a grid of software, hardware and databases, virtual and physical objects and sensors connecting and communicating with each other to serve humanity [5]. IoT aims to create a global network infrastructure to facilitate the easy exchange of commodities, services, and information [6].

The application of IoT in the industry, such as manufacturing and supply chains, is named as Industrial IoT (IIoT) [7]. IoT or IIoT has been applied by some companies to assist in the collection of on-site real-time information, which has successfully improved operating efficiency.

The innovation of IoT benefits companies in the fields related to logistics and affects the operations of enterprises [8]. Moreover, the smarter supply chain has several features like [3]: Instrument, Interconnection, Intelligence, Automation, Integration and Innovation.

According to previous studies, IoT has the following impacts on a supply chain:

1. Warehousing operations: the IoT enables time-saving of joint ordering [9], achieves collaborative warehousing via using smart things. The security and safety level of a supply chain are also enhanced by IoT [10].
2. Inventory: The real-time visibility of the inventory is practicable via the usage of IoT. One of the challenges in traditional SCM was anticipating inventory without real-time visibility. Also, the manual collection of data caused inventory disorder problems. Adding sensors for inventory aims to 100% accuracy rate of inventories.
3. Real-time SCM: Although in traditional SCM information on demand passes only to one partner instead of sharing it, the new technologies of RFID tags enable recording process of all types of information such as production and expiry date, warrant period, product's feature, its' application, and storage condition.
4. Logistic transparency: sensors and networks in the IoT facilitate accurate and timely delivery [11] scanning and recording times are saved by using smart phones [12]. The whole logistic information (transport condition, destination, etc.) is available to the entire supply chain thanks to the smart objects. Consequently, the chance of monitoring and saving goods is indisputably shifted and the cost of returned products is highly declined; besides, the level of customer satisfaction is drastically enhanced. As for IoT application in a supply chain, there are researches been done since 2003; their results are presented on Table 19.1. Studying such researches indicate that no research is organized in the field of IoT application in the supply chain of health care products.

19.2.2 Maturity Levels of IoT

The value of the Internet of Things increases exponentially on a relatively stable and foreseeable maturity curve. What starts with a non-connected device should in theory end in a completely cross-ecosystem automated solution [4].

IoT maturity levels (Fig. 19.1):

1. Dumb: Describes the physical device or product without any connections.
2. Digital: The stage in which the device is digitally enabled.
3. Monitored: Demonstrates remote condition monitoring.
4. Controlled: Presents remote control of a device.
5. Optimized: Includes algorithms, decision-support, and additional services.
6. Autonomous: Relates to self-coordination, automated, and decision-making device.
7. Ecosystem-enabled: The status of smart interaction with other smart objects in the same ecosystem.
8. Cross-ecosystem optimized: Showing smart interaction with devices in other ecosystems.

Table 19.1 The results of literature review

Resource	Results
[13]	The potential benefits of RFID-based investments are able to prevent the rapid depreciation of capital.
[14]	RFID technology is effective in the improvement of systemic mistakes. In addition, its application in taking care of human beings, hospital operations, and patient’s smart medicine basket is considerable.
[15]	The designed models demonstrate the value of exchanging things between special actors in the IoT industry.
[16]	It is possible to locate current human resource and equipment by the active logistic data.
[17]	According to the research of four constructing enterprises, it is concluded that RFID systems have an efficient potential for editing current documents, tracking, and controlling the material.
[18]	The smart system can plan waste collection and act as a guide for operative workers in ordinary and urgent cases.
[19]	IoT application in logistic
[20]	In a food supply chain, smart objects perform the execution, decision making, and learning operations automatically.
[21]	In the conclusion section, a description of a case study and its challenges for holding block chain technologies in future tracking systems of the food supply chain is surveyed.
[3]	They provided a framework for building intelligent, secure and efficient systems. The results showed that the proposed system provides the process of product identification and product tracking globally and reduces time and cost, and then brings customer satisfaction.
[22]	Using BIC technology in industry 4.0 is necessary.

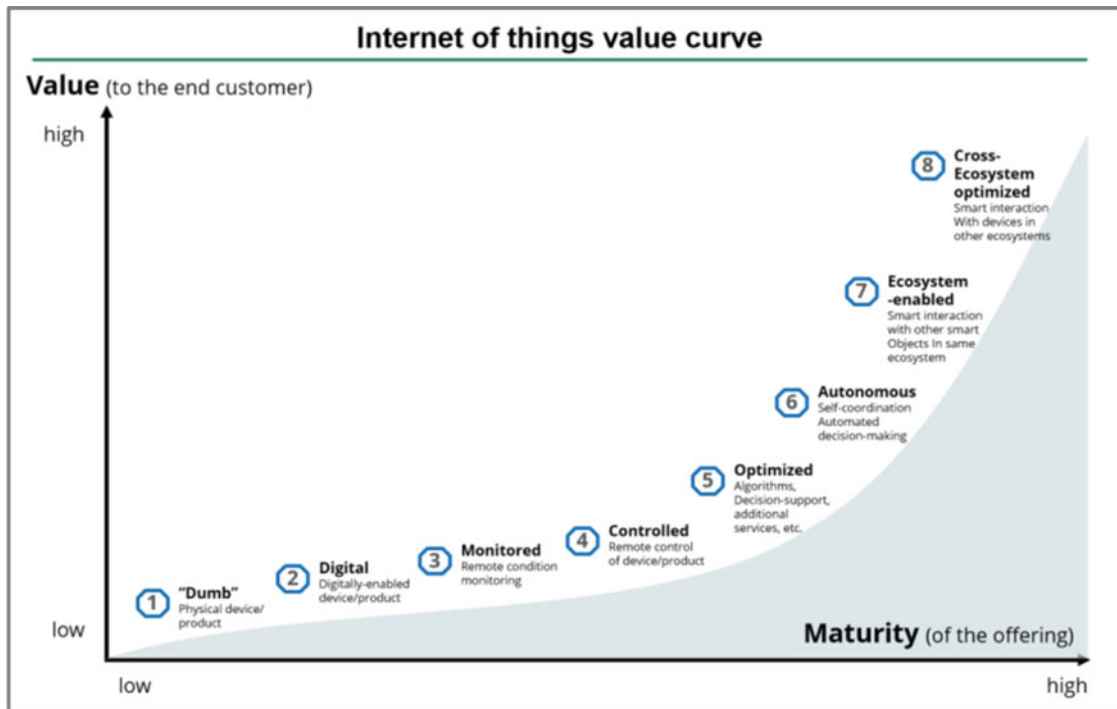


Fig. 19.1 Internet of things value curve [4]

It is essential to consider that the required IoT service should comply with the correspond layer of the maturity level; in other words, each layer of maturity system necessitate special technology and device. When there is an IoT project in a system, requirements of the project’s output should be specified and the proper maturity layer must be defined. In this article, the IoT application in industry considers the required maturity level of each section of the process.

Even though numbers of researches have pointed to the importance of applying the IoT, there is no survey on the maturity of IoT. For example, [13] indicated the importance of using RFID, however, they did not mention points about the feasibility of applying the technology in a special area. In [3], Abdel-Basset et al. discussed smart and secure supply chain in a form of decision making model and calculated security criteria; nonetheless, they did not discuss implementation.

In [22], Luthret et al. rendered the necessity of attention to the IoT in the industry 4.0, again without any consideration on its feasibility and required maturity level. This paper not only surveys the application of IoT in the supply chain of health care products in Iran as a case study, but also pays special attention to the implementation feasibility based on the maturity curve.

19.3 Methodology

This study is practical-oriented; the results are beneficial for the healthcare products' industry. The research approach is deductive-inductive and the strategy is case-study; also, data analysis is hybrid. The procedure of data collection is based on interviews, expert panel, questionnaire, and available documents. The research question is "What are the current challenges of the health products' industry and how to fill these gaps in the light of IoT?" Following steps are applied to answer the question of the study:

First, a comprehensive study on the Internet of Things (IoT), its' properties, and applications in multiple industries is conducted. The accomplishment of this part is identifying the potentials of IoT; however, no survey on the health care products was found. In addition, the IoT maturity curve was studied since accessing the avails of a technology depends on the feasibility of its implementation in the specific country and context; herein, the application of IoT is identified according to the required maturity level. A summary of this study is provided in the literature review. Because of the paucity of information in the industry of distribution and sale of health care products, the field is scrutinized via exploratory studies and expert interviews; moreover, requirements for integrating processes and overcoming challenges are addressed through interviews with experts. In the next section, the practical opportunities established by the IoT potentials are proposed and correspondent maturity layers are assigned to each application according to the maturity curve. The chart of the research process is provided in Fig. 19.2.

19.4 Findings

In this section, IoT applications in the distribution, marketing, sale, and customer service of health products are demonstrated based on researches and expert panel. In the following, outcomes of each section are presented in the form of challenges, IoT-based solutions for the mentioned challenges, and the correspondent maturity layer. It is worth mentioning that questionnaires were designed and provided for the interviewee; after confirming the content, content

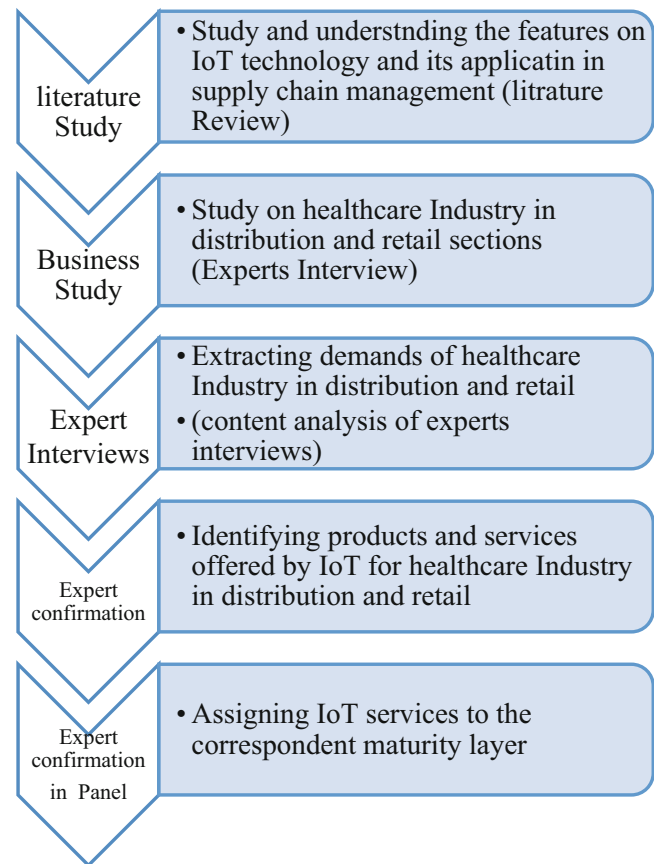


Fig. 19.2 The research process

analysis proceeded by two experts. According to the kappa coefficient, the results were confirmed in terms of reliability. Finally, outcomes were approved by five experts in IoT and five experts in the health care industry. Tables 19.2 and 19.3 illustrate an overview for distribution and sale processes.

19.4.1 Present Processes in Product Distribution (B2B Processes) and Sale (B2C Processes)

The distribution process is defined as the statement below.

"Orders are declared to distributor companies by retailers, ordered products are packed in the warehouse, and the corresponding invoice is then organized. Finally, the mentioned order is delivered to a specific destination." As presented, available processes in distributing health products from receiving ordered forms until delivering them are considered in this paper. The sale process is included by some stages starting with receiving products from distributors to sell them to the end-user.

Details on the stages, their challenges, IoT solutions for challenges and required maturity level are presented in the

Table 19.2 Challenges, IoT based solutions and required maturity level in distribution (B2B processes)

Challenges of distribution process	Potential of IoT to provide the service(Solution)	Maturity Level of IoT to meet services(prioritize of implementation)
1. Store/drugstores send orders individually	A real inventory of shelves is available for distributing companies and there is an automatic ordering process. There will be no need for an employee to visit the store/drugstore.	6th layer
2. The more the products' diversity at the warehouse, the more remarkable mistake arises.	There is a controlling IoT-based gate which compares the ordered and received product	6th layer
3. In some cases, products are missing in the invoice and are not reckoned for payment as the invoices are set manually.	The gate available in the distributor company identifies the tags on products. The invoice will be set based on the identified tags.	6th layer
4. Some of the <i>received</i> cheques are postponed which account for misinformation between the distributor and the retailer.	An IoT based application saves every invoice fee and outstanding amount. Both retailers and distributors have online access to this application.	4th layer
5. As new products arrive frequently in short periods, designing and printing the catalogs would take a long time and demands high cost.	The tags on products help the user to access information about the product which is presented by a smart phone.	3rd layer
6. Training is a permanent activity that should be repeated. Training each retailer continuously is time-consuming and costly.	One session training is sufficient. The information accessed by tags provides complete information.	3rd layer
7. Most of the questions are not asked from the supporting department since retailers are very busy at work.	There is a virtual chatting space to answer immediate questions.	5th layer
8. Although large investment is paid for the advertisement, results are not convincing.	According to the data originated from the IoT application, the best advertisement is presented within social media.	7th layer
9. Long and laborious efforts are required to collect and analyze data to develop the project; nonetheless, data might be also inaccurate.	An accurate report of sold items is available at any time.	8th layer
10. It is difficult to estimate the future demand for a new product launched in the market owing to the fact that there is no feedback on it yet.	A real inventory of products in the warehouse of distributor and shelves of stores and also users' feedback is available in the moment.	6th layer
11. The period time of returning products from the end-user to the distributor is long.	<i>Solution:</i> Returned products are reported in the moment thanks to the smart tag.	4th layer

Table 19.2 for the distribution process and in the Table 19.3 for the sale process.

In the end, to achieve an effective investment, it had better to determine the maturity level of IoT and decide the best solution. Accordingly, the form bellow is proposed and allocated to the applicant companies. Based on the company's situation, a number from 1 to 8 should be assigned according to the IoT maturity level curve:

1. To what extent does the distribution part of your company apply the technology of IoT?
2. To what extent does the sale part of your company apply the technology of IoT?

19.5 Conclusion

Convergence Technology is of paramount importance in the development of industries. One of the technologies in information technology is the BIC which is composed of Big Data, Internet of Things, and Cloud Computing. In fact, the real

application of IoT is practicable if the infrastructure of BIC is prepared. To measure the maturity level of a company, the Technology readiness level in nine levels [23] is used. If the required infrastructure of a company is completely prepared, it stands in level 9. In this case, it is feasible for the company to implement IoT's application with a high maturity level. It should be seriously considered that the maturity level of the IoT in the industry is very essential; every operation that will be equipped with IoT has to be evaluated before any actions whether each layer of the maturity level of IoT can provide a defined value or not. In this paper, a study on the development of distribution and sale supply chain of health care products in Iran is organized according to the demand and gap of information in this specific industry. To perform the survey, challenges of the distribution and sale part of the supply chain of health care products in Iran are first identified by interviews with experts. The supply chain covers four segments; the challenges of all segments are addressed and correspondent solutions for overcoming the challenges are proposed. The items are as follows:

Table 19.3 Challenges, IoT based solutions and required maturity level in sales (B2C processes)

Challenges of sale process	Potential of IoT to provide the service(Solution)	Maturity Level of IoT to meet services(prioritize of implementation)
1. As products are counted manually and their amounts are inserted to computer by an operator, errors may arise. The more diverse the products, the lower precision in analyzing inventories. Ordering requires analyzing the sales' history like records in product storage and shortage. One of the errors that some of the inventory reporting systems make is that out of stocked products are not declared simultaneously.	Shelves and products are equipped by IoT tags. Their instant real inventory is visible. Auto-ordering is done considering decrease in real inventory. The best ordering quantity is calculated by analyzing recorded data and forecasting market potential within cloud computing.	7th layer
2. Sometimes, a professional retailer isn't available at the store/drugstore. Correct and comprehensive information wouldn't be transferred to the user completely. When the end user's demand is unavailable or out of stock, the retailer tries to sell her/him another similar product. Finally, the end-user wouldn't be satisfied with her/his shopping.	Tags stocked on each product include all information of product such as application and ingredients. This information would be available in smart phone. By the presence of the ordering system via IoT, shortage never happens to the inventory.	7th layer
3. A busy queue on cash desk especially in promotion events occurs; the customer spends a long time for payment. In this case, mistakes are possible to occur in providing invoices.	Paying the bills by smart phone without staying in the queue. Whatever collected from shelves is recorded in customer's account and the whole shopping bill will be reduced from costumers' account.	8th layer
4. Analyzing customers' record is a time-consuming process with difficulties in repetition	Managing returned products	6th layer
5. Receiving ordered products from the distributor company (recounting products and checking invoice). Manual recounting products is time-consuming with possible errors.	As customers shopping information and analysis are available in the moment, the process is done very easy and fast.	6th layer

1. The processes of product distribution: 11 challenges and 11 correspondent solutions are identified. Among the solutions, two of them are in the third layer of the maturity level which has a high priority according to their fast resulting characteristic. Two solutions are in the fourth layer, one in the fifth layer, and four solutions are in the sixth layer, one in the seventh and one in the eighth layer.
2. The processes of product sale: five challenges and five correspondent solutions are identified. Among them, two solutions are in the sixth layer, one in the seventh and another is in the eighth level.

According to the mentioned findings and interviews with experts, it is found that the industry of health care products has the infrastructure only for the third layer of maturity level; hence, only two solutions of the third layer are practicable at the moment. To implement other solutions, companies must achieve more maturity in the IoT industry.

It is suggested that the level of IoT technology in companies is measured by the TRL; then awards are dedicated to ones trying to enhance their readiness in this technology.

It is recommended that the government acquire the required infrastructure from universities and knowledge-based enterprises and rent it to the companies. The proposed sug-

gestion can raise customer satisfaction, especially in the COVID-19 pandemic situation.

Herein, it is recommended that research projects are organized and supported for surveying the right solutions and low cost and early accomplishments.

References

1. S. Nižetić, P. Šolić, D.L.D.I. González-de, L. Patrono, Internet of Things (IoT): opportunities, issues and challenges towards a smart and sustainable future. *J. Clean. Prod.* **274**, 122877 (2020)
2. L. Wu, X. Y. A. Jin, D.C. Yen, Smart supply chain management: a review and implications for future research. *Int. J. Logist. Manage* **27**, 395–417 (2016)
3. M. Abdel-Basset et al., Internet of Things (IoT) and its impact on supply chain: a framework for building smart, secure and efficient systems. *Futur. Gener. Comput. Syst.* **86**, 614–628 (2018)
4. K.L. Lueth, IoT Strategy Primer: the new sources of value enabled by the internet of things, IoT Analytics 2015
5. C. Mims. Here's the one thing someone needs to invent before the internet of things can take off. ed: Quartz (2013)
6. X. Liu, Y. Sun, Information flow management of vendor-managed Inventory system in automobile parts inbound logistics based on Internet of Things. *J. Softw.* **6**(7), 1374–1380 (2011)
7. D.X. Li et al., Internet of things in industries: a survey. *IEEE Trans. Indust. Inform.* **10**(4), 2233–2243 (2014)
8. S.J. Grawe, Logistics innovation: a literature-based conceptual framework. *Int. J. Logist. Manage.* **20**(3), 360–377 (2009)

9. R. Angeles, RFID technologies: supply-chain applications and implementation issues. *Inform. Syst. Manage.* **22**, 51–65 (2005)
10. M. Liukkonen, T.-N. Tsai, Toward decentralized intelligence in manufacturing: recent trends in automatic identification of things. *Int. J. Adv. Manuf. Technol.* **87**, 2509–2531 (2016)
11. J. Fang, T. Q. Z. Li, G. Xu, G.Q. Huang, Agent-based gateway operating system for RFID-enabled ubiquitous manufacturing enterprise. *Robot. Comput. Integr. Manuf.* **29**, 222–231 (2013)
12. B. Li, C. Y. S. Huang, Study on supply chain disruption management under service level dependent demand. *J. Netw.* **9**, 1432–1439 (2014)
13. M. Kärkkäinen, Increasing efficiency in the supply chain for short shelf life goods using RFID tagging. *Int. J. Retail Distrib. Manag.* **31**(10), 529–536 (2003)
14. P. Fuhrer, D. Guinard, *Building a Smart Hospital Using RFID Technologies: Use Cases and Implementation* (Department of Informatics-University of Fribourg, Switzerland, 2006)
15. L. Liu, W. Jia, Business model for drug supply chain based on the internet of things. In *Network Infrastructure and Digital Content, 2010 2nd IEEE International Conference on*, pp. 982–986. IEEE, 2010, September
16. L. Atzori, A. Iera, G. Morabito, The internet of things: a survey. *Comput. Netw.* **54**(15), 2787–2805 (2010)
17. Y. El Ghazali, É. Lefebvre, L.A. Lefebvre, The potential of RFID as an enabler of knowledge management and collaboration for the procurement cycle in the construction industry. *J. Technol. Manag. Innov.* **7**(4), 81–102 (2012)
18. L. Zhang, A. Atkins, H. Yu, Knowledge management application of internet of things in construction waste logistics with RFID technology. *Int. J. Comput. Sci. Commun. Technol.* **5**(1), 760–767 (2012)
19. Z.D.R. Gnimpieba, A. Nait-Sidi-Moh, D. Durand, J. Fortin, Using internet of things technologies for a collaborative supply chain: application to tracking of pallets and containers. *Proc. Comput. Sci.* **56**, 550–557 (2015)
20. C.N. Verdouw, J. Wolfert, A.J.M. Beulens, A. Riialand, Virtualization of food supply chains with the internet of things. *J. Food Eng.* **176**, 128–136 (2016)
21. F. Tian, A supply chain traceability system for food safety based on HACCP, blockchain & Internet of things. In *Service Systems and Service Management (ICSSSM), 2017 International Conference on*, pp. 1–6. IEEE, 2017, June
22. S. Luthra, A. Kumar, E.K. Zavadskas, S.K. Mangla, J.A. Garza-Reyes, Industry 4.0 as an enabler of sustainability diffusion in supply chain: an analysis of influential strength of drivers in an emerging economy. *Int. J. Prod. Res.* **58**(5), 1505–1521 (2020)
23. J.C. Mankins, Approaches to strategic research and technology (R&T) analysis and road mapping. *Acta Astronaut.* **51**(1–9), 3–21 (2002)

Sima Marzban, Paul Meade, Marziye Najafi, and Hossein Zare

Abstract

Patient Engagement (PE) promotes the patient's interaction with and contribution to all aspects of care, where patients play an active and informed role in improving healthcare systems, enhancing health outcomes, and avoiding extra-costs, in addition to individual care decisions. Understanding the PE concept is essential for e-health professionals to adopt solutions to interacting intensely with patients. To identify the gaps in stakeholders'— particularly e-health people— worldviews, we conducted a scoping review of the evidence that has been published between 2010–2020. In this review, we included published PE articles that focused on the role of information technology. Our findings showed that stakeholders' solutions have focused primarily on clinical records, communications, education, adherence, and recently, artificial intelligence to optimize the services. The authors focused their attention on the care's aspects regarding cognitive, emotional, economic, behavioral, lifestyle, or wellness dimensions.

Reviewed evidence rarely emphasizes the patients' role in changing organizational policies, care redesign, or healthcare service improvements. We propose a model to

develop PE by multi-stakeholder efforts and interrelated capabilities by coordinating diverse engagement tactics into a seamless orchestration, using versatile Information Technology (IT).

Keywords

Patient engagement · Information technology · Insights · Experience · E-health · Activation · Providers · Payors

20.1 Introduction

Patient engagement (PE) is a growing scheme around the world. Information Technology (IT) solutions provide platforms to engage the patients with their care process, to significantly adapt to situations (such as the current pandemic) in which patients are less likely to visit clinical providers in person. The Medical Institute (IOM) considers access to appropriate information and clinical knowledge as a source of control over individuals' health-related decisions [1]. Furthermore, engaging patients as partners in shared decisions promotes better quality and lower cost, return on investment, and improved outcome measures. Evidently, cost containment, quality improvement, customer retainment, and adherence can explain stakeholders' main drivers for enhancing the patients' role. However, improved health outcomes as the highest level of impact provides the greatest benefits to patients and communities.

Electronic health interventions are known to have great potential to enhance patients' engagement with passive or active involvement strategies.

Providing access to clinical records, text, audio, or video sources for patient education is an example of *passive* engagement. Examples of *active* engagement include mutual relationships like live video communications, virtual pa-

S. Marzban (✉) · P. Meade

University of North Carolina, Gillings School of Global Public Health, SPH Acad Affairs, Key Patient Insights, Chapel Hill, NC, USA
e-mail: Simasi@live.unc.edu; simam@keypatientinsights.com;
Pmeade@email.unc.edu; pmeade@keypatientinsights.com

M. Najafi

Department of Health Economics and Management, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

H. Zare

Johns Hopkins Bloomberg School of Public Health, Department of Health Policy and Management, Department of Global Health Administration, University of Maryland University Global Campus (UMGC), Baltimore, MD, USA
e-mail: hzare1@jhu.edu; Hossein.Zare@faculty.umuc.edu

tient family advocacies, asking about future care or design preferences, and patient contributions to decisions. Using the capacity of IT, our study highlights patient engagement approaches by key actors, developing a multi-stakeholders' conceptual framework to optimize patient engagement, and ultimately patient outcomes.

20.2 Methods

This paper consists of two main sections: a scoping review of available evidence and a multi-actor model fed by our review results.

20.2.1 Scoping Review

This scoping review was conducted in 2020 to identify the gaps in the evidence that are important to healthcare professionals and policymakers. The review provides a picture of the current state of PE approaches accessible by published evidence. Several primary and secondary studies were available to disclose PE understanding through the lens of one of the stakeholders, but a few studies revealed a comparative or collective view of various stakeholders' perspectives on the usage of IT. We conducted the review through Arksey and O'Malley's original framework [2, 3], which has 5 steps, including the following:

20.2.1.1 Research Question

In step one, we identified the research question: How do stakeholders approach PE?

20.2.1.2 Identifying Related Studies

In the second step, we identified related studies. The search was performed in PubMed and Google Scholar databases using the keywords: patient engagement, patient insight, patient involvement, and patient activation.

20.2.1.3 Selecting Articles

The research team included review studies that focused on the above-mentioned keywords from two or more stakeholders' perspectives. Inclusion criteria were: [1] studies written in English, and [2] secondary studies that reviewed PE approaches by two or more stakeholders. The following criteria were used to exclude review studies: non-English language, did not address the study's question, and duplicate studies.

From 4512 published documents, 191 presented thoughts and views on PE related to a primary intervention to engage/activate patients; 22 had a multi-stakeholder intervention, and only 17 were secondary reviews with more than a single-stakeholder perspective on the patients' role. After three authors reviewed the included articles, 17 articles were

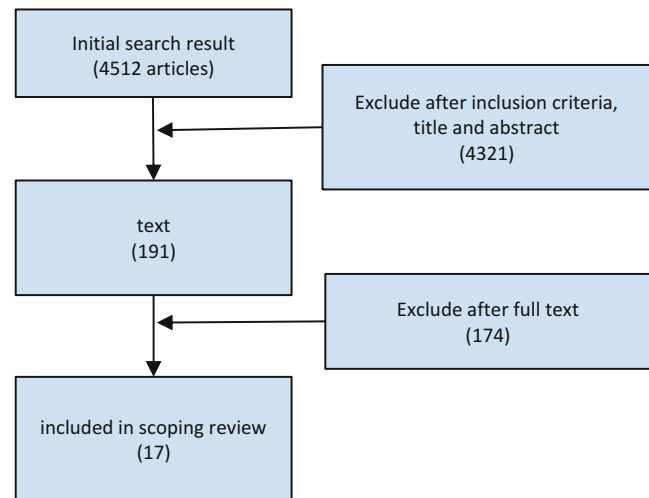


Fig. 20.1 Flowchart of the study selection process

selected for this study. Disagreements were resolved through discussion between the original review authors and in group discussions among authors. From a careful consideration of the selected articles, 17 quality articles were selected for this study. While writing this paper, we looked at the references used by the 17 selected articles when required. The screening process and search results are shown in Fig. 20.1.

20.2.1.4 Tabulating Findings

In the fourth step, data related to understanding and interpreting PE was extracted and tabulated based on the type of stakeholders involved.

20.2.1.5 Reporting Findings

The selected documents were categorized according to the definitions, approaches, focus of PE stakeholders, and examples of engaging methods.

20.2.2 Multi-Actor Model

We developed a new conceptual model that provided a comprehensive view of PE interactions.

20.3 Findings

PE refers to the actions that producer, provider, and patient need to obtain the greatest benefit from the healthcare products and services through the patients' informed involvement. The perspectives of diverse healthcare players that looked into PE were different—a range of benefits from saved costs to averted complications or death. Also, involvement strategies— defined as ranging from entering

data in a form by the patient to consultations for designing the operation room considering aspects are important to patients. Understanding PE from different dimensions of various stakeholder groups plays a crucial role in optimizing the implementation of PE.

20.3.1 Scoping Review Findings

Selected articles showed that PE approaches and interventions could be classified into four areas of interest: E-health solution developers, Insurers (payors), Clinical Providers, and Biotech companies.

20.3.1.1 E-Health Solution Developers

E-health solutions create a digital provider-patient relationship connecting clinical team members, mostly physicians, to patients. Providing access to clinical teams and medical data facilitates communication for real-time decisions. Moreover, software applications and IT platforms allow clinical teams to deliver cooperative and efficient services, impact positively on patient behavior, prevent adverse incidents, and improve health outcomes [1, 4].

Studies show that engaging patients who had physical and mental health problems reported better emotional, physical, and social outcomes [5]. It is essential to consider what are patient expectations, experiences, and operational preferences from eHealth interventions and the IT design. Otherwise, while providers work strictly with those boards, patients ignore the utilization and its ultimate advantages [6]. A competitive landscape of patient engagement among digital market navigators resulted in beneficial groups' simultaneous efforts to add a type of PE digital tool to their services. Patients being prescribed a new drug may be asked to enroll in their health care provider's (HCP's) hospital or clinic electronic portal at the same time in the life science company's patient engagement program, or in a wellness application from their health insurance company. The situation leads to a potentially overwhelming and conflicting patient experience despite the best intentions of all stakeholders. Though potentially complex, there is the opportunity to coordinate these diverse engagement tactics into a seamless orchestration of touchpoints with the unified purpose of supporting optimal health outcomes [7].

Methods to Engage Patients by E-Health Solution Developers

Investing in digital patient interaction tools warrants a return on investment in the health industries [8]. E-health interventions focused mainly on visit timing, clinical records, and patient education [9]. A variety of innovative models also engage inpatients, including before hospitalization navigators, visit and consultation administrators, video patient

educators, billing transparency processors, and during hospitalization audiovisual options for medical purposes (such as pain management through distractions). Those initiatives can provide patients and clinical teams with advanced support mechanisms rather than general health information and one-way communications [1].

As long as providers recognize the need and arrange for mutual communications and feedback, IT can actively engage patients with the clinical process [10]. Transforming care delivery to the optimal extent of expected patient engagement will be achievable if IT professionals consider the influencing factors related to crucial stakeholders such as patients, providers, payors, and the biotech industry.

20.3.1.2 Insurers (Payors)

Payors see PE as an interaction policy with patients and families who suffer chronic and costly medical conditions to maximize the value of paid expenses, so-called value-based care, and payments. In fact, value-based payment can adjust the risk-sharing between stakeholders, the strategy that significantly affects how organizations develop physicians' networks, invest in service lines, and plan for clinical locations and treatment programs [11]. Insurance organizations encourage stakeholders to consider the quality and cost of services provided. To this end, paying targets for value should improve the health system's performance, which is certainly not possible with traditional payment methods [12]. Physicians' long-term efforts have proven to provide quality and valuable services following the financial results [13].

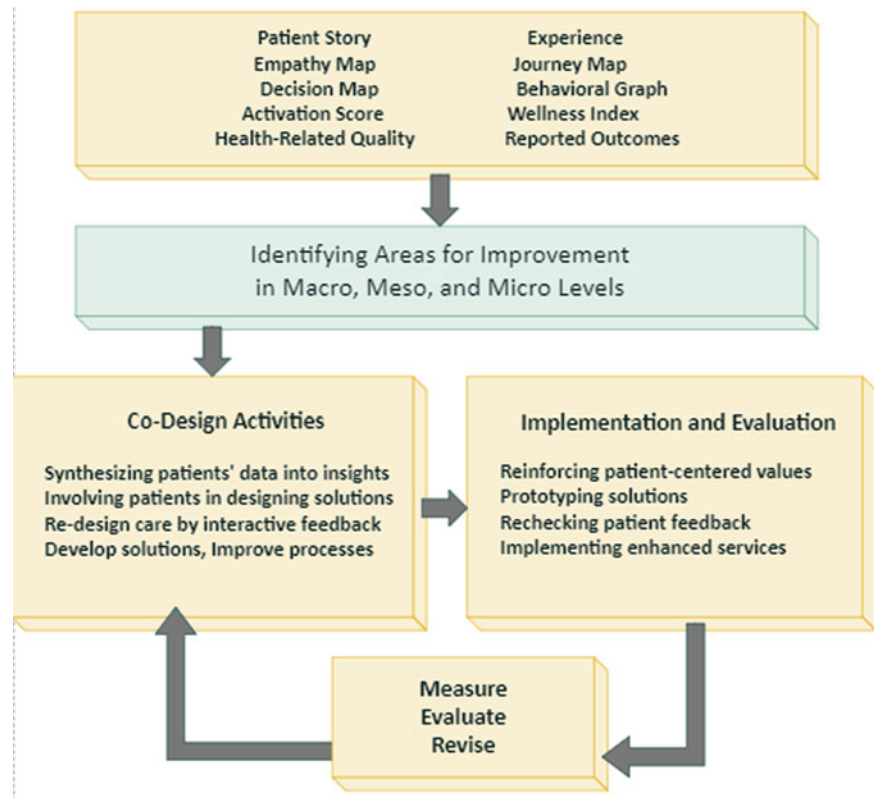
Methods to Engage Patients by Insurers (Payors)

United Healthcare, a large national payor, emphasizes providing health information to patients to reduce the information gap between patients and providers. This includes sharing detailed information about selecting providers, supporting treatment decisions, and planning diagnostic tests and treatments. There is also a "payback" program to Accountable Care Organizations (ACOs) that provide more quality and low costs through preventive health programs that offer positive financial incentives by reducing premiums to employees with healthy habits [8]. The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) is a survey used as a basis for the patient's experience of care in a value-based purchasing program by insurers. Hospitals required report data for this plan to receive their full inpatient prospective payment (IPPS) annual payment update (APU). They may be penalized if they have not met the required quality [14].

20.3.1.3 Clinical Providers

For a convincing and meaningful interaction with patients, providers must consider consumer relationship language,

Fig. 20.2 Integrated patient-side data capturing diagram, insights to impacts package. (From Key Patient Insights (KPI) website, by Marzban, 2020, www.keypatientinsights.com)



communication timing, smooth flow of information, and clear responsibilities. Evaluation of patient engagement by satisfaction studies and outcome measures will also help providers enhance the well-engaged patients' retainment in their network; however, engaged patients are different from satisfied patients [15]. PE should be operated and acknowledged at three levels:

micro-level (medical approach), meso (institutional), and macro-level (system). Micro-level such as educating patients for self-assessment purposes is usually considered as a PE achievement while patients are often not assigned to be engaged in macro-level like the redesign of an organizational policy or procedure [9]. The level of patient interaction seems to affect the results of efforts that are dedicated to patients and family engagement [16]. To this end, having a strong culture of interaction between patients, providers, and healthcare professionals is required by six principles of partnership: learning, empowerment, transparency, responsiveness, and respect [17].

Methods to Engage Patients by Clinical Providers

A range of mechanisms have been applied by providers based mostly on information provision, patient activation, and patient-provider collaborations [17]. They could be classified due to the degree of participation: low-level participation such as one-way consulting advice, and high-level interactions such as joint design (co-design) or partnership

strategies [16]. Co-design as the most mature type of engagement optimizes providers' insights on actual patient-side data and can help providers understand PE through moving from insights to achievable impacts (See Fig. 20.2) [18].

20.3.1.4 Biotech Companies

Historically, the pharmaceutical industry has focused on developing science and medicine to prevent or treat diseases. Patient-centric involvement means involving the patient in the drug development journey from discovery to the marketplace. Leading health-care companies have involved the patients and families, while some large pharmaceutical enterprises have lagged behind and are still focused on traditional markets [19].

Pharmaceutical companies are inherently seeking to make a remarkable profit, which should be noted that the acquisition and use of this profit must be consistent with the claim of patient centrality. There are several challenges for pharmaceutical companies in the PE area: doubts about commercial success, lack of standard process, insufficient sharing of pharmaceutical industry experiences in patient engagement, lack of research on patient centrality, overuse of available information sources (such as physicians and Medical Science Liaisons (MSL)), the notion that direct industry interaction with patients is inappropriate or unauthorized, and conflicts of interest (the idea that the industry should not engage with patients). Most of these challenges are solved by sharing experiences and learning [20].

Engaging steps: Here are the main steps to engage bio-pharma companies with patients:

- Changing the mindset – turning the patient focus into a vision throughout the company leaders and staff.
- Driving cooperation – patients and families and other stakeholders’ participation to examine patient-centered solutions.
- Learning and sharing – Challenging the situation, recording experiences, learning, and sharing lessons [20].

20.3.2 The Multi-Stakeholder PE Model: Role of Information Technology

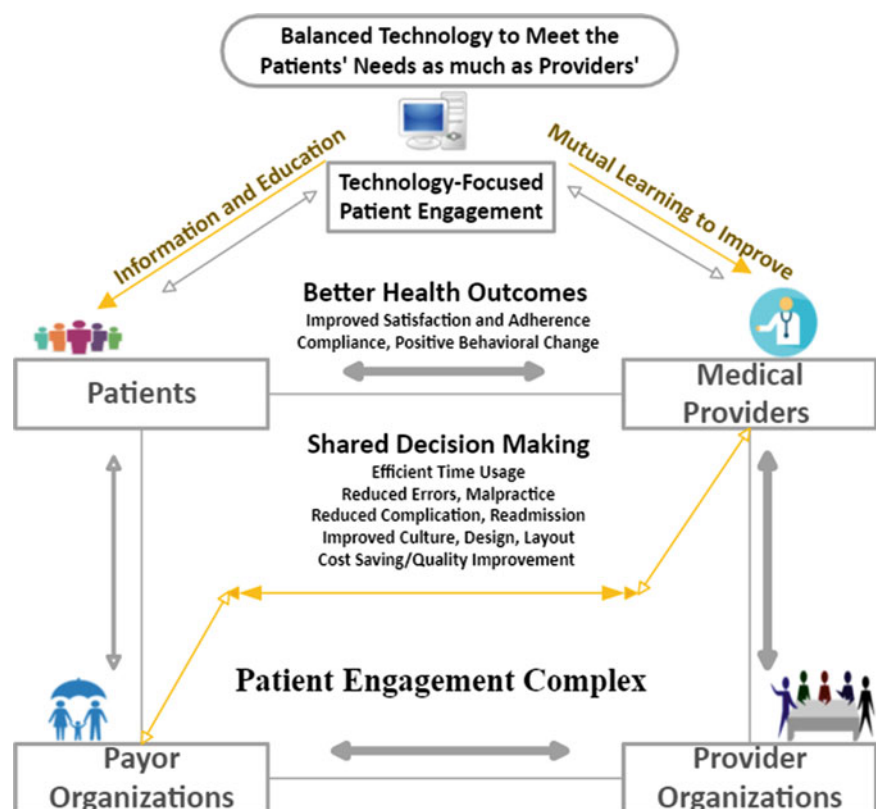
By combining stakeholders’ perspectives, we developed a conceptual framework (Fig. 20.3) that is particularly advisable to apply PE by means of various actors’ determination: Data and Technology partners, Medical Providers and Delivery roles (physicians and nurses), Healthcare Organization (System), Patients/Families, and Payor Organizations.

20.3.2.1 Technology-Focused Patient Engagement

IT and data play essential roles in engaging patients with the services and clinical teams. IT platforms should provide user-friendly, interactive, on-time, and easily understandable

information for diverse users. Solutions are more competitive that meet the needs for information in real-time, not only about the disease but also about the resources of social support and self-care. Addressing providers’ need to transform, IT is evolving to make the best of artificial intelligence, machine learning, and data analytics. However, developing the capacity to inform providers and learn from patient-driven insights seems to be a profound level of achievement. The technology puzzle’s challenging part is how market players navigate the patient engagement to the stakeholders’ priorities, including patients and physicians who are most likely to develop their humanized clinical inter-actions. Another crucial question is that: how patients should encounter and interact with various changing applications and software over time and insurance coverage changes (particularly in the US) to play an active role in their health services. Is it a time to provide an integrative and sustainable digital environment for the patient’s lifelong access? If yes, how will payors and providers customize their connection to the constant patients’ digital rooms through inter-operable technologies when insurance or provider plans switch? The service model will change for the better and add to the efficiency that already exists through consideration of how the different actors look at PE and the recently emerged technologies. While solution developers pay attention to providers’ expectations and preferences, it is crucial to meet the patients’ priorities. For instance, the billing process is vital to providers and

Fig. 20.3 Conceptual framework to engage patients using capacity of information technology



payors, but financial support is essential for the patients. Consumption of a prescription shows adherence and sustained revenues for pharmaceuticals, but perceived quality of life changes during a drug or device usage are important to patients. IT solutions might keep an eye open to how patients interact with the software or apps and how their concerns, such as the need for customized information and financial or emotional support, are met.

20.3.2.2 Medical Providers

Individual clinical roles are influenced by personal approaches, inherent characters, and internalized behaviors. Wellness-oriented attitudes by clinicians and clinical team members often establish promotive health discussions. The provider's mutual communicative style also develops a shared discussion process that is more appreciative than assertive orders to patient values. Adequate knowledge and practice facilitate sharing the right information with understandable language and managing the discussion process with a two-way, calm, and kind context. There are time-related, cultural, non-verbal, audio, and visual considerations to help patients remain actively involved and responsible for their health efforts and outcomes. Providing adequate incentives and inputs for decisions, using audio and visual enabler aids may support patients to actively engage with shared decisions.

20.3.2.3 Healthcare Organization (System)

Healthcare Organizations affect both patients and medical providers directly and indirectly. They ought to give PE a strategic priority as a dynamic and evolutionary process requiring multi-stakeholder interactions aligned with the patients' preferences, activation, empowerment, and optimal use of the engagement. Providers emphasize PE values and patient-centered leadership style in their organizations to steer subordinate systems to process goals such as efficient time usage, fewer errors, better design/culture, shared decision-making, adherence, and compliance. At the same time, they secure target outcomes such as better health status, disease relief, loyalty, and retention. By considering community-based care and social determinants of health (SDH), providers can transform how they interact with their patients. Top leadership commitment to a patient-centered culture and investments in solutions to maximize end-users' engagement might be reflected in written organizational policies, rules, measures, structural roles, evaluations, and continuous training and improvements.

20.3.2.4 Patients

Patients are at the core of the model—a fully engaged patient stems from a series of personal traits and interpersonal behaviors. A patient who prefers to take an active role in the clinical team and intentionally builds knowledge and confidence would be a better decision partner than a passive

patient. Heard, valued, and informed patients take accountability for the healing process and see their health outcomes as within their control. Engaged patients seek information and resources to change their behavior and do not hesitate to express the need for assistance or support. Empowered patients continuously ask for solutions when faced with stress/pain points and demand transparent billing/pricing information. The most mature type of engagement occurs when the patients participate in improvements and contribute to the care redesign.

20.3.2.5 Payor Organizations

If payors understand the real value and meaning of engagement, they may see PE as a business imperative. Some experts criticize current value-based payment because of inconsistencies with evidence-based medicine, transparent payment policies, and as misleading about spending less without changing practice path and harming patients. Capturing patient-side data such as patient experiences, patient journey maps, and wellness indexes generate value for payors by saving unnecessary services and wasteful expenditures. Gathering and analyzing customer feedback, if done through payors, producers, or providers, is usually tainted with internal bias; however, it is always possible to work with trusted vendors who indirectly communicate with users to provide transparent insights. The consistency of payment policies with evidence-driven treatment saves payors' money because of fewer errors, readmissions, and follow-up visits. In contrast, when physicians are concerned about being compliant with insurance protocols that are not scientifically proven, prolonged treatment complications and costs would burden the payor organization.

20.4 Conclusions

Based on the review, existing PE approaches are heterogeneous, patchy, and unpolished. PE strategies are also not always consistent with patients' values and benefits as expected from the term, but PE often watches over the desired advantages of digital technologies, provider individuals, healthcare organizations, and most of all financial organizations to optimize trust, consumption, and adherence. Even patients' organizations and groups face gaps in their basic presumptions of PE. In the absence of powerful patient associations, there is confusion about whether PE means to act like a better consumer and care-receiver or to excel in the user roles as counselors and co-designers for the healthcare products and services for which they pay high costs. PE is a dynamic and evolutionary process that requires multi-stakeholder interactions. The synergy leads to the healthcare processes and healthcare outcome improvements. The proposed systems-based model enables healthcare professionals to systematically lead and develop PE activities within and exterior to

their organizations. Although the proposed systems-based model is not expected to employ a unified PE framework, it will help all stage players be aware of other influencers' impact on what happens to patients' contributions.

References

1. J.E. Prey, J. Woollen, L. Wilcox, A.D. Sackeim, G. Hripcsak, S. Bakken, et al., Patient engagement in the inpatient setting: a systematic review. *J. Am. Med. Inform. Assoc.* **21**(4), 742–750 (2014)
2. H. Arksey, L. O'Malley, Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* **8**, 19–32 (2005)
3. R. Armstrong, B.J. Hall, J. Doyle, E. Waters, 'Scoping the scope' of a Cochrane review. *J. Public Health* **33**(1), 147–150 (2011)
4. S. Sawesi, M. Rashrash, K. Phalakornkule, J.S. Carpenter, J.F. Jones, The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. *JMIR Med. Inform.* **4**(1), e1 (2016)
5. S.L. Ivey, S.M. Shortell, H.P. Rodriguez, Y.E. Wang, Patient engagement in ACO practices and patient-reported outcomes among adults with co-occurring chronic disease and mental health conditions. *Med. Care* **56**(7), 551–556 (2018)
6. S. Triberti, S. Barello, The quest for engaging AmI: patient engagement and experience design tools to promote effective assisted living. *J. Biomed. Inform.* **63**, 150–156 (2016)
7. Why Great Technology Is Not Enough: 5 steps to scaling white-glove service and support. Steven Huddleston, President & CEO, PELITAS – Wednesday, December 23rd, 202. Access date: 1/4/2021. Available at: <https://www.beckershospitalreview.com/why-great-technology-is-not-enough-5-steps-to-scaling-white-glove-service-and-support.html>
8. L.G. Sandy, R.V. Tuckson, S.L.J.H.A. Stevens, UnitedHealthcare experience illustrates how payors can enable patient engagement. **32**(8), 1440–1445 (2013)
9. P. Rieckmann, A. Boyko, D. Centonze, I. Elovaara, G. Giovannoni, E. Havrdová, et al., Achieving patient engagement in multiple sclerosis: a perspective from the multiple sclerosis in the 21st Century Steering Group. *Mult. Scler. Relat. Disord.* **4**(3), 202–218 (2015)
10. S. Barello, S. Triberti, G. Graffigna, C. Libreri, S. Serino, J. Hibbard, et al., eHealth for patient engagement: a systematic review. *Front. Psychol.* **6**, 2013 (2016)
11. J. Bynum, V. Lewis, Value-based payments and inaccurate risk adjustment—who is harmed? *JAMA Intern. Med.* **178**(11), 1507–1508 (2018)
12. J. Lynn, A. McKethan, A.K. Jha, Value-based payments require valuing what matters to patients. *JAMA* **314**(14), 1445–1446 (2015)
13. J.A. Hirsch, T.M. Leslie-Mazwi, G.N. Nicola, M. Bhargavan-Chatfield, D.J. Seidenwurm, E. Silva, et al., PQRS and the MACRA: value-based payments have moved from concept to reality. *Am. J. Neuroradiol.* **37**(12), 2195–2200 (2016)
14. Seibert SA, Stroud A, Cassel L, Huebner C. Improved HCAHP Scores and a DEU Culture of Excellence. <https://www.hcahpsonline.org> Centers for Medicare & Medicaid Services, Baltimore, MD. Access Date: March 15, 2021
15. M. Bellows, K. Kovacs Burns, K. Jackson, B. Surgeoner, J. Gallivan, Meaningful and effective patient engagement: what matters most to stakeholders. *Patient Exp. J.* **2**(1), 18–28 (2015)
16. Y. Bombard, G.R. Baker, E. Orlando, C. Fancott, P. Bhatia, S. Casalino, et al., Engaging patients to improve quality of care: a systematic review. *Implement. Sci.* **13**(1), 98.16 (2018) Ontario's Patient Engagement Framework. Health Quality Ontario. 2017. p. 8. ISBN 978-1-4606-9801-3
17. S.W. Grande, M.J. Faber, M.A. Durand, R. Thompson, G. Elwyn, A classification model of patient engagement methods and assessment of their feasibility in real-world settings. *Patient Educ. Couns.* **95**(2), 281–287 (2014). <https://doi.org/10.1016/j.pec.2014.01.016>
18. Integrated Patient-Side Data Capturing Diagram: KPI, Insights to Impacts Package. Available at: www.keypatientinsights.com
19. M. Lush, D. Rosner, C. Zant, S.N. Tuthill, Patient engagement strategies in a digital environment, Life sciences companies respond to changing patient expectations. Deloitte Review. Issue 18 (2016). <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-18/patient-engagement-strategies-changing-patient-expectations.html>
20. D. du Plessis, J.-K. Sake, K. Halling, J. Morgan, A. Georgieva, N. Bertelsen, Patient centricity and pharmaceutical companies: is it feasible? *Ther. Innov. Regul. Sci.* **51**(4), 460–467 (2017)

Juliana Damasio Oliveira, João A. L. de Moraes Junior, and Rafael H. Bordini

Abstract

We conducted a mapping study to investigate how the scientific community uses Ambient intelligence technologies to assist visually impaired people. Our initial search identified a total of 807 publications; after applying our selection criteria, we accepted 65 publications. We seek to show which technologies, methodologies, techniques, architectures, features, and evaluations are most used. Results indicated that only 15.53% have a specific focus on people who are visually impaired. Most of the results were published in 2015. Most authors used their own architectures. Many results used techniques such as Detection and recognition, Artificial Intelligence, Networking, and others.

Keywords

Ambient intelligence · AAL · Visual impairment · Mapping study · Home care · Home environment · Smart environments · Smart spaces · Smart home · Visual disability

21.1 Introduction

Ambient Intelligence (AmI) refers to sensitive, adaptive, proactive physical environments that respond to people and objects' actions and cater to people's needs [1]. Researchers believe that AmI is a vision of intelligent computing, where people will be supported by the environment they inhabit [2]. Based on this vision, Ambient Assisted Living (AAL) tech-

J. D. Oliveira (✉) · J. A. L. de Moraes Junior · R. H. Bordini
School of Technology, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil
e-mail: juliana.damasio@acad.pucrs.br; rafael.bordini@pucrs.br

nologies have emerged to assist, monitor, and promote a caring environment, especially for people who have special needs. The group of special needs includes people who are visually impaired, to whom this work is particularly addressed.

According to the World Health Organization (WHO), globally, there were approximately 285 million people who are visually impaired (PVI), of which 39 million were blind [3]. These people face needs and barriers that are different from elderly people. They need to understand the state of their own home and have difficulty in managing their usual tasks [4, 5], for instance.

We conducted a Mapping Study (MS) that is a type of systematic literature review that provides an overview of a topic or research area [6]. This MS has the objective of assessing how AAL is used for the benefit of PVI. We classify existing solutions according to the techniques, technologies, architectures, methods, and evaluation performed.

The paper is structured as follows. Section 21.2 shows the protocol used in the MS. Section 21.3 presents the results obtained. Section 21.4 presents a discussion on the results and shows the conclusion about MS.

21.2 Materials and Methods

We use the MS protocol proposed by Petersen et al. [6]. The main goal of this study is to identify how AmI technologies are used to assist people who are visually impaired. We divided the MS into three main parts: Plan, Conduction, and Report.

Plan: Based on the purpose of this MS, we defined 7 research questions: RQ1—What technologies were used? RQ2—What architectures were used? RQ3—What techniques were used? RQ4—What methodologies adopted to design the systems? RQ5—What features are provided? RQ6—How were these environments/technologies evalu-

Table 21.1 Search expression

Keyword	Alternative term and synonym
Visual impairment	(blind OR visually impaired OR visual impairment OR visual disability OR blindness OR unsighted OR low vision OR disabled people) AND
Ambient intelligence	(ambient assisted living OR ambient intelligence OR smart home OR smart Care OR smart service OR smart homecare OR ambient-intelligence environment OR smart environments OR home environment OR smart spaces OR home care)

ated? RQ7—What are the challenges and limitations of these environments/technologies?

We choose 5 relevant digital libraries in Computing science and Healthcare: ACM Digital Library¹, ScienceDirect², IEEEExplore³, Scopus⁴, and Pubmed⁵. Afterwards, we identified the keywords related to the research topic, such as “visual impairment” and “ambient intelligence”, as well as their alternative terms and synonyms. Although this MS is addressed to find results focused on visual impairment, we added “disabled people” in the string because this is a general term that includes several disabilities, including visual impairment. We combined these terms using logical operators to create the search expressions shown in Table 21.1. The search expressions were adapted according to each digital library’s particularities, to no alter the intended meaning.

We defined a primary study as a control article for the validation of search expression. This article was previously identified in non-systematic searches. The article title is “RUDO: A Home Ambient Intelligence System for Blind People” [7]. If this article was in the digital libraries, they had to come in the search with the search expression that we created. If the control paper is not returned during the search, the string needs to be adjusted until they do so. We created some selection criteria for the selection of publications:

Inclusion (I1) Result containing in the title, in the keywords, or in the abstract some relation with the theme of this review (Ambient intelligence and Visual impairment).

Exclusion (E1) Not published in English; (E2) Similar or duplicate results, only the most recent will be considered; (E3) Results that are not related to the theme of this work (Ambient intelligence or Visual impairment); (E4) Books and abstracts from conference presentations; (E5) Narrative reviews, comparative studies, surveys, and other systematic reviews; (E6) Studies set in environments other than home

environment; (E7) Studies set in fields other than computer science or engineering; and, (E8) Results prior to 2009. We used a cutoff date because we wanted the latest technology.

Conduct In this phase, we execute the protocol defined previously. After applying the search string to each digital library and downloading the results in Feb 2019, we imported the results on Start,⁶ a tool to help classify the publications. We also used the Kappa Method for Measurement of inter-rater reliability [8], for better alignment between researchers and reduced bias. Using the Start tool, we performed the selection criteria for inclusion and exclusion of publications, in two phases.

First, one of the authors of this paper read just abstract, title, and keywords of the papers in the search results. We assigned the status of “accepted” to the papers that met the inclusion criteria. These papers were selected to be fully read later; In the second phase, these results were reviewed by two authors, individually. Then, we compared the results from the two researchers, and when it was different, they discussed until reach a consensus. In this phase, some studies were initially considered appropriate for inclusion, but after being fully read, they were excluded.

Report The search expression application in each digital library brought a total of 807 papers, of which 129 were duplicates. After applying the selection criteria, in the first phase, we selected 127 papers. Afterwards, in the second phase, we selected 65 papers⁷ after applying the selection criteria by reading the entire paper. We applied the Kappa method and obtained 92.85% of agreement. According to Cohen’s Kappa Interpretation, this is an “Almost Perfect agreement” result [8].

We identified that 84.61% of the results actually implemented solutions for AAL, while 15.38% of the results contained non-functional prototypes [9–14], frameworks [11, 15, 16], design methodologies [14], case studies [17], and techniques analysis [18].

Among the type of users, 60% of results combined some types of users, for example, elderly and disabled people, or elderly and blind users. In this case, we consider each type individually for analysis. The majority of the results are directed toward elderly people (35.92%). A considerable part of the publications deal with visual impairment in general, focusing on “disabled people” (33.98%). Only 15.53% of the publications explicitly cite the focus on people who are visually impaired. 14.56% of publications show approaches to other types of users, such as people who are physically disabled (e.g., have paralyzed limbs or hearing impairment)

¹<https://dl.acm.org/>

²<https://www.sciencedirect.com/>

³<https://ieeexplore.ieee.org/Xplore/home.jsp>

⁴<https://www.scopus.com/>

⁵<https://www.ncbi.nlm.nih.gov/pubmed/>

⁶http://lapes.dc.ufscar.br/tools/start_tool.

⁷The list of accepted paper references is available at <https://github.com/julianadamasio/mappingstudy.git>.

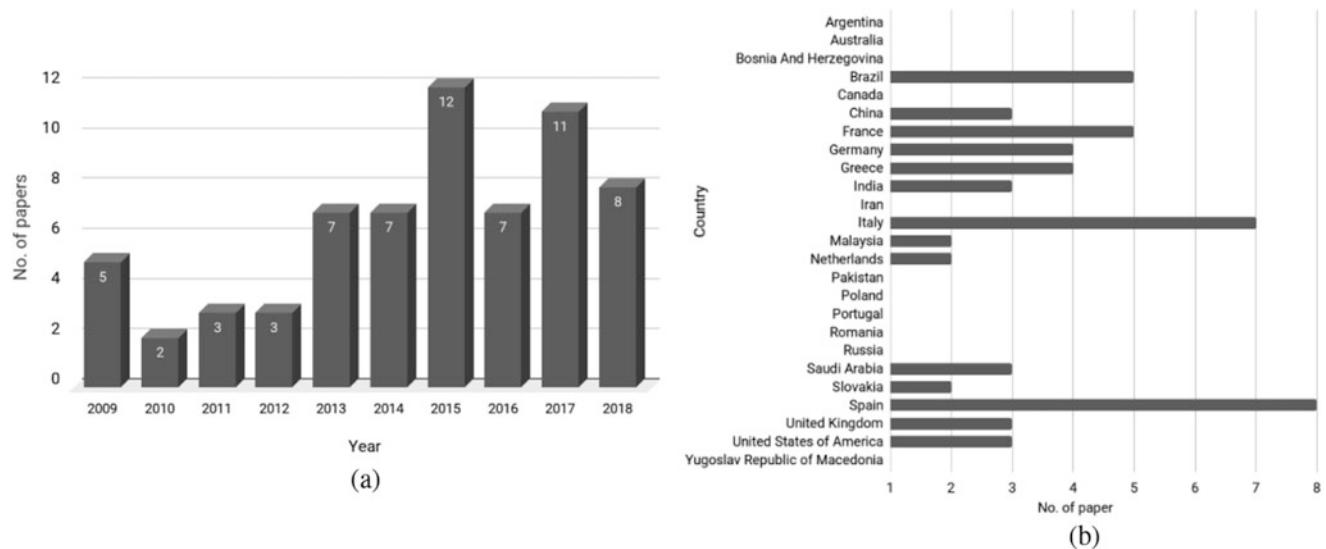


Fig. 21.1 Number of papers per year and country. (a) Papers per year. (b) Papers per country

and families of the elderly and disabled people. We will refer to publications from here as follows:

- general papers: papers with generic focus, for example, for disabled people.
- specific papers: papers that contains a specific study to PVI.

In general papers, Fig. 21.1a shows the distribution of papers per year, being 2015 and 2017 the years with the highest concentration of papers. Figure 21.1b shows that Spain (8), Italy (7), Brazil, and France (5) were the countries that most published on ambient intelligence. We found 40 publications from the Europe, followed by Asia (13 papers) and America (9 papers). The number of papers per country/continent was collected according to country of the first author institutional affiliation. Considering specific papers, 2015 (31.25%) and 2018 (25%) were the years with the highest concentration of papers. The countries that most published were Brazil (3) and Italy (3). The Europe (9 papers) has the most publications for this type of user as well.

21.3 Results

Referring to question **RQ1**, we identified 94 different technologies in general papers. We considered each technology individually in analysis. The most used technologies were: Sensors (6.89%), Computer/Laptop (4.31%), Arduino (3.44%), Camera (3.44%), Environmental sensor (3.44%), RFID (3.44%), Smart device (3.44%). There are a variety of technologies, and the vast majority of publications did use a

combination of technologies. Some results had a combination of RFID + Sensors, or Motion sensors + Cameras, or Devices + Sensors, for example. Considering specific papers, most used technologies were: Computer/laptop (7.84%), Accelerometer (3.92%), Arduino (3.92%), Cameras (3.92%), Microphone (3.92%), Motion sensor (3.92%), Raspberry Pi (3.92%), and Smart device (3.92%). These publications also used a combination of technologies.

Addressing to question **RQ2** in general papers, most architectures used were the authors' own architectures (39.28%), Service-Oriented Architecture (SOA) (10.71%), Multi-agent systems (MAS) (7.14%), and Others (e.g., ROS, multi-robot, OSGi, KNX) (7.14%); 35.73% of results did not declare the architecture they used. The specific papers also used own architectures (56.25%), SOA (12.5%), and MAS (6.25%); 25% did not declare which architecture was used.

Referring to question **RQ3**, we found results that make use of a large number of techniques, so we group these techniques as follows:

- Artificial intelligence: encompasses computational vision, decision tree, machine learning, deep learning, reasoning engines, neural network, agents, and Bayesian network.
- Detection and recognition: include algorithms for recognition and detection of voice, gesture, activity, movement, face, finger, pose, eye blink, motion, human, and speech.
- Networking: encompasses the internet of things, communication paradigms and protocols, synchronisation, web services, and data sharing.
- Others: include a navigation system, object recognition, Robot Operation System, Visual-Range Odometry, among others.

Several results used combined techniques, and we considered each one separately for analysis. The most used techniques in general papers are Detection and recognition (34.95%), Artificial Intelligence (22.33%), Others (24.27%), Networking (14.56%), and not declared (3.88%). Considering specific papers, the most used techniques were: Others (36%), Detection and recognition (28%), Artificial Intelligence (24%), Networking (8%), and not declared (4%).

Referring to question **RQ4** in general papers, most results (68.42%) did not adopt a particular methodology for the AAL design solutions. Some results have mixed some methodologies. The most used methodologies that they reported were: Service-oriented (10.52%), User oriented (8.77%), Agent oriented (8.77%), and Design science research (DSR) (3.50%). When considering the specific papers, most publications do not state which methodologies they used (61.11%). The most used methodologies they reported were: User oriented (16.66%), DSR (11.11%), Agent oriented (5.55%), and Service oriented (5.55%).

About the question **RQ5** in general papers, the features most cited were Environmental/appliance/device control (5.36%), Environmental/device/people monitor (5.36%), Health monitoring (4.02%), Turning on/off the devices/sensors (4.02%), Fall detection (2.68%), Detection of dangerous situations (2.01%), and Prevention of medical emergencies (2.01%). Several results have more than one feature proposed. For example, in [19] there are control and monitoring of devices and turning on/off the devices. Among the specific papers, the features most cited were: warning whenever it detects a potential hazard (6.66%), detection and recognizing objects (6.66%), turning on/off the devices/sensors (6.66%), and avoid obstacles (6.66%).

Referring to question **RQ6**, we categorised the solutions evaluations as follows:

Type of environment: when the authors evaluated the proposed solution in:

- controlled environment: experiments that occurred in research laboratories, controlled by the researchers.
- virtual environment: experiments in which they tested algorithms, dataset, simulations of real environments in virtual environments.
- real environment: when the proposed solution was installed and evaluated in real environments such as homes for the elderly, homes for the PVI.

Type of users: when the authors evaluated the proposed solution with:

- end-user: the proposed solution was evaluated by users who are the focus of the solution, such as the elderly, PVI.
- user: the proposed solution was evaluated by general users who are not the focus of the solution.

- not evaluated by the user: end-users and other general users did not evaluate the proposed solution. This type usually occurs when the authors evaluated the solution in a virtual environment.

Among the general papers, 36.36% not declared the type of environment used, 27.27% did not declare the type of user, and 3.64% of the results did not evaluate the solution yet. Most of the results (29.09%) were not evaluated by any user, 25.45% were evaluated by end-users, and 14.55% were evaluated by other users. Regarding the type of environment used in the evaluation of solutions, they used: virtual environment (25.45%), a controlled environment (18.18%), and a real environment (12.73%). Some results evaluated two types of environment: a controlled environment and a real environment (1.82%), and virtual environment and a controlled environment (1.82%).

About the specific papers, the most common type of environment is controlled environment (18.75%), real environment (18.75%), virtual environment (6.25%), and controlled environment and real environment (6.25%). 50% did not declare the environment. Most of the evaluations (37.5%) were performed with end-users, 18.75% with users, and 6.25% without users. Additionally, 37.5% did not declare the type of users that performed the evaluation.

Referring to question **RQ7** in general papers, 66.66% did not declare challenges and limitations. The rest of the results could not be quantified; hence, it is not possible to provide any percentage of the results. There are limitations on the use of Kinect, such as software and hardware [20], and maintaining a certain level of hand steadiness to control [21]. Some papers reported solution implementation just in a single-user context [7, 22]. Moreover, problems related to real-time were also identified as to track five persons simultaneously [23] and perform some analysis and exploitations [24]. Also, some limitations of occlusion [20], recognition [25, 26], and performance [27].

The limitations and challenges of specific papers were related to implementation because they were implemented just in a single user context [7, 22]. 75% of the results did not declare their limitations and challenges.

21.4 Discussion and Conclusion

We started this review with the aim of identifying studies for visually impaired people in the field of ambient intelligence. During the research, we found that most studies focus on disabled people or combine the types of users. Only 15.53% specifically mentioned people who are visually impaired. These findings demonstrate that more studies need to be carried out for these users.

Answering our research question “to identify how AmI technologies are used to assist people who are visually impaired”, the literature has shown that the authors have combined technologies to implement interesting features such as warning whenever it detects a potential hazard, detection and recognising objects, turning on/off the devices/sensors, and avoid obstacles. They also preferred to use their own architectures or SOA. In many studies, there was no participation of end-users during the system development and evaluation. Many human-computer interaction methods demonstrate the importance of involving the system end-user in the design phases, such as interaction design method [28].

We believe that systems have to be developed for and by the end-user. In this sense, we agree with [29]; if the end-user is not involved in the system design, likely, the technology will not be accepted by them. Thus, researchers should consider the specific interests, demands, and needs of the end-users, especially when it comes to disabled people. Despite this, we demonstrate an advance in this topic by showing the features, technologies, techniques, architectures, and methodologies that are being used by the systems developed for AmI, both for the general public and people who are visually impaired.

Acknowledgments This study was financed in part by CAPES—Finance Code 001.

References

1. E. Aarts, R. Wichert, Ambient intelligence, in *Technology Guide* (Springer, New York, 2009)
2. F. Sadri, Ambient intelligence: a survey. *ACM Comput. Surv.* **43**(4), 36:1–36:66 (2011)
3. W.H. Organization et al., Universal eye health: a global action plan 2014–2019 (2013)
4. M. Vacher, B. Lecouteux, J.S. Romero, M. Ajili, F. Portet, S. Rossato, Speech and speaker recognition for home automation: Preliminary results, in *International Conference on Speech Technology and Human-Computer Dialogue* (2015), pp. 1–10
5. J.D. Oliveira, R.H. Bordini, A survey on the needs of visually impaired users and requirements for a virtual assistant in ambient assisted living, in *16th International Conference on Information Technology-New Generations (ITNG 2019)* (Springer, New York, 2019), pp. 449–453
6. K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering, in *International Conference on Evaluation and Assessment in Software Engineering*. (BCS Learning & Development Ltd., Swindon, 2008), pp. 68–77
7. M. Hudec, Z. Smutny, Rudo: a home ambient intelligence system for blind people. *Sensors* **17**(8), 45 (2017)
8. M.L. McHugh, Interrater reliability: the kappa statistic. *Biochem. Med.* **22**(3), 276–282 (2012)
9. K. Lefevre, S. Totzauer, A. Bischof, A. Kurze, M. Storz, L. Ullmann, A. Berger, Loaded dice: exploring the design space of connected devices with blind and visually impaired people, in *Nordic Conference on Human-Computer Interaction* (2016), p. 31
10. S. Lim, L. Chung, O. Han, J.-H. Kim, An interactive cyber-physical system (cps) for people with disability and frail elderly people, in *International Conference on Ubiquitous Information Management and Communication* (2011), p. 113
11. C.V. Gándara, C.G. Bauza, Intellihome: a framework for the development of ambient assisted living applications based in low-cost technology, in *Latin American Conference on Human Computer Interaction* (2015), p. 18
12. M. Cunha, H. Fuks, Ambleds collaborative healthcare for aal systems, in *IEEE International Conference on Computer Supported Cooperative Work in Design* (2015), pp. 626–631
13. S.K. Nayak, N.S. Chavan, N. Srinath, User centered inclusive design for assistive technology, in *IEEE Annual India Conference* (2016), pp. 1–6
14. R. Picking, A. Robinet, V. Grout, J. McGinn, A. Roy, S. Ellis, D. Oram, A case study using a methodological approach to developing user interfaces for elderly and disabled people. *Comput. J.* **53**(6), 842–859 (2009)
15. S. Zolfaghari, R. Zall, M.R. Keyvanpour, Sonar: smart ontology activity recognition framework to fulfill semantic web in smart homes, in *International Conference on Web Research* (2016), pp. 139–144
16. E.Z. Tragos, M. Foti, M. Surligas, G. Lambropoulos, S. Pournaras, S. Papadakis, V. Angelakis, An IoT based intelligent building management system for ambient assisted living, in *IEEE International Conference on Communication Workshop* (2015), pp. 246–252
17. M. Danancher, J.-J. Lesage, L. Litz, G. Faraut, Online location tracking of a single inhabitant based on a state estimator, in *IEEE International Conference on Systems, Man, and Cybernetics* (2013), pp. 391–396
18. S. Banitaan, M. Azzeh, A.B. Nassif, User movement prediction: the contribution of machine learning techniques, in *IEEE International Conference on Machine Learning and Applications* (2016), pp. 571–575
19. A. Hussein, M. Adda, M. Atieh, W. Fahs, Smart home design for disabled people based on neural networks. *Proc. Comput. Sci.* **37**, 117–126 (2014)
20. B. Yao, H. Hagra, D. Alghazzawi, M.J. Alhaddad, A big bang–big crunch type-2 fuzzy logic system for machine-vision-based event detection and summarization in real-world ambient-assisted living. *IEEE Trans. Fuzzy Syst.* **24**(6), 1307–1319 (2016)
21. L.E. Anido, S.M. Valladares, M.J. Fernandez-Iglesias, C. Rivas, M. Gomez, Adapted interfaces and interactive electronic devices for the smart home, in *International Conference on Computer Science & Education* (2013), pp. 472–477
22. M. Hudec, Z. Smutny, Advanced scene recognition system for blind people in household: The use of notification sounds in spatial and social context of blind people, in *International Conference on Computer Science and Application Engineering* (2018), pp. 159:1–159:5
23. J. Han, J. Han, RGB-D human identification and tracking in a smart environment, in *Computer Vision and Machine Learning with RGB-D Sensors* (Springer, New York, 2014)
24. A. Larab, E. Conchon, R. Bastide, N. Singer, A sustainable software architecture for home care monitoring applications, in *IEEE International Conference on Digital Ecosystems and Technologies* (2012), pp. 1–6
25. M.E. Abidi, A.L. Asnawi, N.F. Azmin, A. Jusoh, S.N. Ibrahim, H.A.M. Ramlil, N.A. Malek, Development of voice control and home security for smart home automation, in *International Conference on Computer and Communication Engineering* (2018), pp. 1–6
26. C. Xu, W. Li, J.T.C. Tan, Z. Chen, H. Zhang, F. Duan, Developing an identity recognition low-cost home service robot based on

- turtlebot and ROS, in *Chinese Control And Decision Conference*, (2017), pp. 4043–4048
27. B. Pontes, M. Cunha, R. Pinho, H. Fuks, Human-sensing: low resolution thermal array sensor data classification of location-based postures, in *International Conference on Distributed, Ambient, and Pervasive Interactions* (2017), pp. 444–457
28. Y. Rogers, H. Sharp, J. Preece, *Design de Interação* (Bookman Editora, Porto Alegre, 2013)
29. M. Choraś, S. D’Antonio, G. Iannello, A. Jedlitschka, R. Kozik, K. Miesenberger, L. Vollero, A. Wołoszczuk, Innovative solutions for totally blind people inclusion, in *Ambient Assisted Living* (CRC Press, Boca Raton, 2015)

Voice for the Voiceless: Developing a Low-Cost Open-Source Communication Device for the Speech Impaired

Travis Smith and Vasilios Pappademetriou

Abstract

Advancements in the affordability and availability of Internet of Things (IoT) devices have led to incredible innovations in the field of speech-generating devices. The introduction of devices such as tablets, cellphones, and mobile computers has allowed individuals with speech disorders to have a medium in which to easily communicate without purchasing expensive and specialized medical equipment. Even though these devices increased access to this technology, it is still out of reach for many. Whether it is the cost of the device, cost of the speech generating software, or access to a reliable internet connection, this technology is inaccessible to some of the people who need it most.

The focus of this project is bringing this voice technology to the oppressed and disadvantaged by creating a fully open-source device assembled with off the shelf parts that is a fraction of the cost of similar alternatives. The goal was to accomplish this with minimal compromises to quality and usability. This was achieved with a Raspberry Pi computer, touch screen, battery pack, and a plastic casing. It overall met the quality and usability expectations of an Alternative Communication (AAC) device for around \$150 USD and shows that traditionally expensive AAC equipment can be made more accessible to people, without compromising usability. It hopefully will motivate others to research areas where off the shelf parts and open-source software can be used to increase

the accessibility of otherwise expensive and specialized technologies to benefit the lives of others.

Keywords

AAC-Devices · Speech Assistant · Internet of Things · Open-Source · Text to Speech · Medical-Devices · Python · Raspberry Pi · Autism Spectrum Disorder · Speech Generating Devices

22.1 Introduction

Speech Generating Devices (SGD's) have been around for the past several decades. They have allowed people to have a voice of their own, who otherwise would not have had the means to. Advances in computer and internet technologies have allowed these devices to become more accessible and usable to those with speech, voice, language, and hearing disorders. These devices come in several variants depending on the type of help needed, and within these categories, come many subcategories depending on the severity of assistance required. Overall, they allow people to communicate in a more meaningful and effective way and participate in conversations in their day to day lives.

There are three main assistive device categories; the first is Assistive listening devices (ALD's). The simplest example of this kind of device is a hearing aid, which filters and amplifies the sound around the user to allow them to hear more effectively. The second is Augmentative and Alternative Communication Devices (AAC's), which is targeted by this research. AAC devices allow users with speech disorders to express themselves effectively. They can range from a simple keyboard-based device to devices with touch interfacing, text to speech transcriptions, or even image to speech translation. The third category is alerting devices, these get

T. Smith (✉)

Cyber Forensics and Security, Department of Information Technology and Management, Illinois Institute of Technology, Chicago, IL, USA
e-mail: tsmith41@hawk.iit.edu

V. Pappademetriou

Department of Information Technology and Management, Illinois Institute of Technology, Chicago, IL, USA
e-mail: vpappade@iit.edu

the attention of individuals by notifying them that some kind of event is taking place. Alerting devices are implemented through a visual queue or even a loud sound [13].

22.2 Three Types of AAC-Devices

Depending on the individual, there are some circumstances where different AAC device types will be more successful than others [14]. This section outlines the three main kinds of AAC devices and how each can be useful to help facilitate speech.

22.2.1 Unaided AAC-Devices

Augmentative and Alternative Communication Devices (AAC's) are the main focus of the research. They allow people with congenital and acquired disabilities to enhance their speech. "Those who may benefit from AAC devices are individuals who have autism spectrum disorder, cerebral palsy, developmental disabilities, developmental apraxia of speech, or other genetic disorders" [2].

Depending on the type of disability, a different form of AAC device may be necessary for that individual, so AAC devices tend to come in one of three forms. Unaided AAC, Low tech AAC, and High Tech AAC [2]. Unaided AAC "does not involve the use of any external devices or materials. Examples of unaided AAC include manual signs and natural gestures (e.g., head shake for "yes" and "no" waving to say "hello" or "goodbye")" [17].

Unaided AAC is an excellent option for those who would rather not use an external electronic device to communicate or those who cannot afford the cost of low tech or high tech AAC devices. However, some people may not be able to communicate effectively using Unaided AAC devices, and could require some form of external electronic assistance, making cost a severely important factor.

22.2.2 Low-Tech AAC-Devices

Low tech devices are part of the aided forms of AAC. These devices usually are not battery powered and are typically more affordable to the consumer. An example of a popular low tech AAC device is the Picture Exchange Communication System [3]. This device was developed using a method that "avoids difficulties inherent in other systems by requiring very few prerequisites; in fact, the only prerequisite is that the individual can indicate (e.g., by reaching for an item) what he or she wants, in a way that can be shaped into exchanging a physical symbol such as a picture" [4].

Most of the advantages of Low Tech AAC devices stem from the lack of battery-powered components. This in turn makes the devices cheaper to manufacture, more accessible to people without plentiful access to electricity, and tends to be more robust and can be used in environments where high-tech devices cannot. These environments can be harsh weather conditions, near or in water, or where power may be scarce.

Some disadvantages of Low Tech AAC devices include the following. Owing to being either non-powered or function with very low power, the vocabulary selection available to the user may be limited at a given time. Words or pictures may not be added without physical effort from either the user themselves or an outside source. Another consideration is that communication relies on the user showing another person their selection, instead of audibly communicating. This reliance may be restrictive since the user may want to get someone's attention but cannot do so without other methods. Third, due to the device's physical nature, it may be bulky and restrictive to carry around if the user wants to have a large vocabulary selection at all times. Extra pictures and word choices may need to be carried around and swapped out. They may also get lost over time.

22.2.3 High-Tech AAC-Devices

High tech devices are also an aided form of AAC and allow the user to most closely imitate the experience of verbal communication. They are battery-powered and normally use touch-screen interfaces and digital voices to convert text or images to speech. These devices can run sophisticated software vocabularies with multiple voices that are fully accessible and can be thoroughly customized. The operating systems of these devices can be Windows, Android, or iOS [9].

The benefits of high-tech devices are as follows. First, due to being a computing device, a vast amount of vocabulary can be programmed into AAC applications. Second, this vocabulary audibly is played for the user and the individual(s) they are communicating with. This playback is closer to the natural flow of communication between people. Third, these devices can allow users to not be dependent on others to communicate. For example, person with a low-tech AAC device may require a partner to help them switch out vocabulary lists, pictures, etc. These high-tech devices can be customized to each individual's needs.

Of course, these high-tech devices also have their disadvantages. They require battery power to function, which means the user must have access to electricity to charge the device. Also, some of the high tech AAC devices may have a learning curve for some users, especially those who used

legacy systems. Regarding children using AAC devices, “the main reason given for lack of use (37%) was that the child was still learning to use the AAC system and therefore was not ready to use it at home” [12]. With today’s children growing up using technology, such as tablet devices, a potential hypothesis could be that children are more likely to use their AAC devices at home, due to their ease of use compared to legacy systems.

Some high tech AAC devices also require an internet connection to function. While this requirement may allow for more advanced human-sounding digital voices, it can prohibit some people from obtaining and using the technology if they cannot afford internet access or do not have it as an option at all due to location. Third, these high-tech devices are universally more expensive than the low tech and unaided tech options, where “commercially available AAC devices range from low-tech single-function products with prices above 100 USD to more advanced high-tech multi-functional products with prices exceeding 15,000 USD” [17]. Ideally, there should be a device that offers high tech devices’ features and convenience for the same cost as low-tech counterparts. That was the goal and the mission of this research.

22.3 Motivation Behind the AAC System

22.3.1 ASD Prevalence

One type of disorder, which affects speech is Autism Spectrum Disorder (ASD). Of these individuals with ASD, it is “estimated that 30% of them will fail to develop vocal output capabilities” [17]. This is a large number of people when scaled to the total population of countries, and the world, with studies in the United States indicating that the “estimated ASD prevalence was 2.47% among US children and adolescents in 2014-2016” [16]. A similar prevalence across the world was found, with studies in South Korea showing that the rate was as “high as 2.6 % per 10,000 (95 % CI 1.9–3.4 %)” [8].

Outside of the US and South Korea, ASD prevalence seems to fluctuate. For example, the “reported prevalence of ASD in South Asia ranged from 0.09% in India to 1.07% in Sri Lanka, which indicates that up to one in 93 children have ASD in this region. An alarmingly high prevalence (3%) was reported in Dhaka city” [7]. There could be different reasons for this difference in prevalence, such as the sample size of the studies. In the South Asia study, “sample sizes ranged from 374 in Sri Lanka to 18,480 in India. The age range [in the study] varied between 1 and 30 years” [7]. While there is a difference in the prevalence of ASD worldwide, it appears that a large number of people have ASD, and thus, a subset of people may need help with their speech. It is also essential

to keep in mind that ASD is only one type of disorder that may benefit from AAC devices.

Not only are AAC devices useful for individuals with ASD, but other cutting-edge devices can be used for assistance, specifically with visual conceptualization and reward reinforcement. Augmented Reality can be used to “inculcate skills as well as particular desired behaviors. It can also facilitate reinforcement and favorable responses in the effective execution of a small task [15]. Augmented reality can now be used with a smartphone, showcasing further that technology is increasing the reach of medical and therapeutic devices and treatments.

22.3.2 iPad and AAC Applications

The release of the iPad in 2010 revolutionized the AAC market. Its touch screen was large enough to fit many pictures and words on it at once, it was easy to use, and much more affordable than dedicated AAC devices. At its release in 2010, and even into 2011, the iPad “prices range from \$499 for the 16GB Wi-Fi model to \$829 for the 64GB Wi-Fi + 3G model” [11]. However, even at this more affordable cost, even the iPad was very much out to reach for most, especially considering that some AAC applications for the iPad can cost hundreds of dollars. In addition to the high cost, users are left at the mercy of the software developers to keep that application up to date with new iOS releases and iPad hardware revisions. At times, developers will abandon their application, and it will be made unusable on newer operating system versions. Apple has introduced rules and regulations in regard to outdated applications, where they implemented “an ongoing process of evaluating apps, removing apps that no longer function as intended, don’t follow current review guidelines, or are outdated” [1].

These application updates are also not limited to operating system compatibility. If the AAC community requests new features to the application, it is up to the discretion of the developers to implement and push those changes to the App Store.

22.3.3 Android and AAC Applications

As for Android AAC device alternatives, there are many reasons why purchasing a cheap Android device and using free AAC software is not the best solution. First, “there is a huge number of Android versions in use on a wide variety of hardware” [10]. Depending on the operating system, Android application developers need to specifically target their applications to a version of Android and an Android device while developing the application. At times, updating

an Android application to support a new Android version requires a new software development kit (SDK), which can break some of the application features. This requires the developers to modify the application to support the newest SDK version.

Thus, due to the many types of Android devices, all running potentially different versions of the Android operating system, downloading a free or paid AAC application puts the user at the developer's mercy for updates and support for their specific device, and if the application is free, those updates have a greater chance of never occurring. These free AAC applications can also be filled with spyware, adware, and sometimes even malware to profit from running on consumers' devices. These attributes lead to an extremely poor user experience, at times not even up to par with the worst iOS alternatives. All applications need to be approved by Apple before being released on the App Store, while Google tends to have more leniency for play store applications.

22.3.4 Design Requirements and Motivations

During the AAC device's design process, the idea of giving people more access to the life-changing technology while simultaneously providing the same functionality of competing high-tech AAC devices was essential. Developing a device that was able to be expanded upon by the AAC community was also a priority. That way, users would not have to rely on a single entity to maintain and update the software. With this goal in mind, the following design requirements were created.

First, the device must be low cost to manufacture and not be sold at a significant profit. At the time of writing, a new iPad from Apple can be purchased from \$329 [1]. The goal for the AAC device was to keep the total cost under \$150, including both hardware and software. The second requirement was to keep the device portable, similar to a tablet so that users can take it with them without much difficulty. Third, the device must be assembled from off the shelf parts; that way, it can be repaired, maintained, and upgraded without having to purchase an entirely new device. Fourth, the AAC software for the device must be open source and publicly accessible. This way, the AAC software can be customized to fit an individual's needs, and the AAC community can maintain it over time. This prevents developers from abandoning their AAC applications and then releasing a new version that must be purchased. This software will be open to modifications and free to download and use.

Finally, the device must be able to work without an internet connection, requiring that the text to speech processing is done locally. This local processing proved to have advantages and disadvantages. However, it would make the device more accessible to people without internet access.

22.4 Hardware Choices

22.4.1 Main Compute Device

To meet the design requirements, intelligent hardware choices with the best cost to reliability ratio were imperative. For the computing device, the Raspberry Pi Model 3 A+ was chosen for its balance of power and portability. This model of Raspberry Pi weighs 29 grams and has a 64-bit quad-core processor that runs at 1.4GHz, 512 MB of RAM, Dual-band 802.11 ac wireless, Bluetooth capabilities, and Audio/Video output for \$30 USD. This device can run a full operating system, with all the required features needed to run the AAC application. It also can function with only 5 V/2.5A DC power via a micro-USB connector, meaning that a slimline battery can power it.

22.4.2 Power Delivery

The power delivery system for the device was chosen with portability, longevity, and maintainability in mind. After investigating, the Maker Focus Raspberry Pi battery and USB hub was chosen. This battery has a capacity of 3800mAh and plugs into the USB hub, which then is connected to the Raspberry Pi to deliver power. What is great about this system is that the battery is replaceable and upgradable, as is the USB hub. Preliminary testing has shown that the Raspberry Pi running the AAC application can be powered for around 4 hours using this specific battery; however, a smarter operating system and software configuration could increase this timeframe. The battery costs \$25 USD, making it an affordable choice for the project as well.

22.4.3 User Interaction

The user interface system was a critical part of the hardware decision-making process. Ideally, the user should be able to reach and read all parts of the interface without strain on their motor skills or eyes. The interface also needed to be touch-based, as that most closely resembles the iPad's interface system and makes interaction intuitive. Owing to these requirements, the Raspberry Pi 7" Touch Screen Display was chosen. This hardware component was the most expensive, costing \$60 USD. However, it is arguably the most important as it is what the end-user will be directly interfacing with to communicate. There were cheaper alternatives available; however, these touch screens were much smaller and did not offer the same kind of functionality and user experience. The touch experience on an iPad is exceptionally refined, so imitating that experience as much as possible was important.

The device casing was also a difficult design decision. Ideally, lightweight and low-cost plastic cases for the device could be manufactured using 3D printing. However, considering that the 3D printing market is still developing, and creating this device with off the shelf parts was a requirement, the SmartPi Touch 2 case was chosen. While thicker than an iPad and other conventional tablets, this case offers a great compromise of portability and protection, making it ideal in the event of being dropped. The case comes with an optional stand, which may prove useful in stationary environments, such as on a desk in a classroom. This case costs \$27 USD, which is more than a 3D printed case; however, it delivers the functionality needed for the device and meets the design requirements.

Overall, the device's hardware cost \$142 USD. This price is less than half of the cheapest iPad option available today in 2020 and will include in that price the open-source, free software. Later in the paper, future hardware and software decisions will be discussed as to how this device can be refined to get this price even lower, which would make it even more accessible to those in need.

22.5 Software Choices

The software choices of the project are arguably just as important as the hardware. The following kinds of software were considered during the design process, the operating system, the programming language, and the library choices within that programming language to create the AAC application.

22.5.1 Operating System

The operating system was predestined to be some Linux flavor, as the Raspberry Pi kernel is open source and closely follows the main Linux kernel by Linus Torvalds. Raspberry Pi also has an official operating system, called Raspbian OS. It is Debian based, licensed using free and open-source software licenses, and is officially supported by the Raspberry Pi foundation. Since this operating system was so closely connected to the Raspberry Pi ecosystem, it was ultimately chosen to be the operating system to use in the AAC device.

22.5.2 Programming Language

The programming language was chosen based on a few factors. The first concern was finding a widely supported language used by others on the Raspberry Pi. After investigating, the Python programming language was the most popular.

Python also comes bundled with the Raspbian operating system. There were many Python tutorials available at the Raspberry Pi project's website, which included directions on creating many types of Internet of Things (IoT) devices using Python and the Raspberry Pi. Overall, Python's support in the Raspberry Pi ecosystem seemed to be unmatched with other programming languages.

22.5.3 TTS and GUI Libraries

Looking further into the language using the lens of AAC applications, two major features were considered. The first feature considered was the available Text to Speech libraries for the language. The second was the graphical user interface libraries, and how well they would function on the Raspberry Pi.

The text to speech library choice was vital, as it needed to have the capability to work offline to meet design requirements. This option would supersede the choice of programming language if needed. Luckily, Python has a text to speech library named `pyttsx3`, which works offline. The library also functions across operating systems and has support for Linux, Windows, and MacOS. It accomplishes this by using the operating system's built-in speech synthesizer software. For Linux, this software is `Espeak`, which is licensed under the GPLv3 license, making it ideal to meet the requirements of the AAC application.

With regard to the graphical user interface (GUI), the following considerations were made. First was how intensive the GUI was on the hardware. While powerful, the Raspberry Pi is limited in its hardware capabilities, and since this AAC device will be battery-powered, hardware utilization will be essential to make sure the user can get ample time out of their device before needing to recharge.

The main GUI toolkits considered were Tkinter, wxPython, PyGTK, and PyQt. These toolkits all have their own advantages and disadvantages. In the scope of AAC applications, which should be simple and easy to use through a touch interface, two of these toolkits were heavily analyzed for this use case, PyQt and Tkinter. PyQt is robust and an advanced GUI framework for Python. It allows users to create advanced and clean interfaces that look like professional desktop applications. Tkinter, on the other hand, is much more straightforward. It allows users to create simple GUIs with basic functionality, such as creating new tabs, buttons, screens, etc. It ideally would not be used for a professional desktop application.

AAC applications by nature should be designed with ease of use in mind and should not have widgets and complicated screens that could confuse the user. The AAC application should allow the user to organize words and pictures into categories, and then have those items fill the screen to tap

to allow the device to speak for them. Thus, due to this simplicity, Tkinter was chosen as the GUI library for the AAC application.

22.6 Device Assembly

22.6.1 OS Installation and Hardware Assembly

Assembling the device was reasonably straightforward, thanks to the off the shelf nature and compatibility of the components chosen. The Raspberry Pi 3 A+ came with everything needed to get started. The first task was to install Raspbian OS on the Raspberry Pi using the included 16GB microSD card. While the Raspbian image could have been downloaded and then formatted to the SD card using a tool like Rufus, the Raspberry Pi foundation has created a tool to install the operating system easily. This tool is NOOBS (New Out of The Box Software).

NOOBS contains Raspbian and will install it for the user. It also includes a list of other supported operating systems for the Raspberry Pi, which can be selected and installed. After the installation of Raspbian OS, the device was installed into the case. This process was simple, as the case was designed around the Raspberry Pi itself.

The next step was to install the 7" touchscreen onto the case and connect it to the Raspberry Pi. This process was completed using the included adaptor board, ribbon cable, and jumper connectors. The ribbon cable connected to the DSI port of the Raspberry Pi, and the jumper wires connected to the GPIO pins. The adaptor board handles the power and data signal conversions from the screen to the Raspberry Pi.

After the touchscreen was installed, the device's final piece was the battery pack and USB hub. The battery plugged into the USB hub board, and connected to the Raspberry Pi's USB power input, using a mini-USB cable.

Overall, the assembly process took less than 20 minutes the first time around, and most certainly could be done in under 10 minutes with previous experience. This attribute is extremely important since the device can be manufactured, repaired, and upgraded quickly, with very few components. Once assembled, the Raspberry Pi was powered on and was tested to make sure that all components were working properly, and that the battery could deliver sufficient power to the device (Fig. 22.1).

22.7 Developing the Application

22.7.1 Backend Speech Functionality

With the programming language and corresponding libraries chosen for the device, it was time to begin developing the

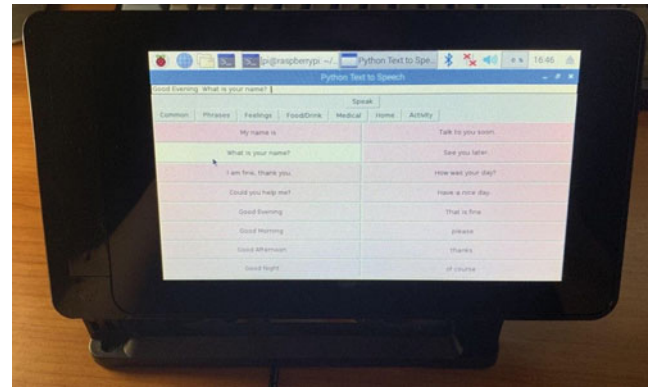


Fig. 22.1 Frontal view of the device

AAC application. Before working on the GUI design, the backend text to speech functionality was implemented. This process included downloading and importing the pyttsx3 library and testing the various voice functions, such as the speaking rate, volume, and voice type.

22.7.2 GUI Design

After an appropriate voice profile was created, GUI design could begin. The goal was to mirror the use and feel of the paid AAC applications, so the GUI's of these applications became a reference during the design process. As this was the first rendition of the device, a simple layout that consisted of several pages of words placed in a grid type formation was chosen. Each page represented a different speech category, such as everyday words and phrases. The categories are feelings, words, phrases related to food and drink, medical, words related to the home, and activity phrases. Of course, the forward vision is that these categories will be configurable, where the user can create their pages with ease and customize the words on each page, and support replacing these words with images.

22.7.3 Application Logic

To start, a generic speech class was created. This class gets passed to each page and includes the logic for aggregating the chosen words and phrases and prepares them to be spoken by pyttsx3. It also includes the GUI logic for creating the word input box and the button to speak the aggregated words. Next, all of the other pages were created with their respective methods. This functionality allowed for flexibility between the pages, as specific vocabulary lists can be created for each page. Each page class contained the logic for displaying the page, the vocabulary lists, and the ability to turn each word in the vocabulary list into an interactive button the user can

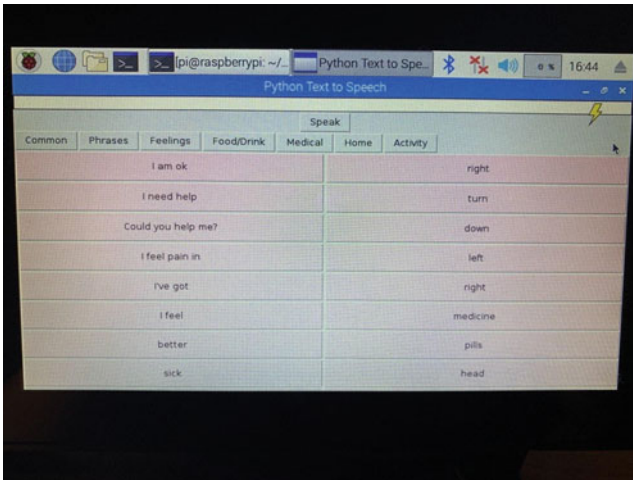


Fig. 22.2 AAC application GUI up close

press to append the word to the sentence builder, which is located at the top of the GUI.

At the top of the GUI is a text box where the selected words and phrases queue up to form a sentence; it also allows the user to type in their own words to be spoken. Once the word or phrase has been constructed, the user taps on the speak button, and the backend `pyttsx3` library handles converting those words in the sentence builder into speech.

Finally, the `MainView` class was created. This class contains the methods for initializing all of the GUI components, including the pages and the container in which the pages were placed. The entry point to the application simply sets up Tkinter, calls the `MainView` methods to initialize all of the components, and then begins the main loop, allowing the GUI to run continuously. For ease of use, a bash script was created to run the AAC application when the device starts and boots into Raspbian (Fig. 22.2).

22.8 Device and Application Strengths

22.8.1 Cost and Ease of Assembly

As a proof-of-concept AAC application utilizing low cost, off-the-shelf parts, and open-source software, this device meets most of the hardware and software requirements, and with a bit of tuning, it can certainly be a competitor against the paid, high-end AAC applications on the market.

The greatest strength of the project is the total cost of \$142. This price includes the Raspberry Pi, case, battery pack, USB hub board, and 7" touch screen assembly. Outside of AAC devices, it is by itself a fully functional personal computer capable of running a full operating system. The release of Raspberry Pi's was revolutionary to the low-cost

internet of things community and has helped bring computing around the world.

Even with the low-price, the components are also all off the shelf and can easily be purchased online or at retail stores worldwide and can be assembled easily by anyone without tools. This easy assembly is especially important, as it decreases the barrier of entry even further for this device.

22.8.2 Device Customization

Another strength is that each AAC device can easily be customized to fit the needs of the user. Different Raspberry Pi models can be used as the base compute unit if the user needs an even cheaper device. The Raspberry Pi Zero, which costs \$5 USD, has been tested and can run the AAC application. Smaller touchscreens can also be configured if the user needs a more portable device. The Raspberry Pi Zero W screen, at just 2.8 inches, can easily fit inside a pocket like a cell phone.

Further research and development will be done to test out an AAC device made from the Raspberry Pi zero, this smaller touchscreen, and a smaller battery pack device to decrease the cost even further and increase the device's portability. These changes will be discussed later but shows that there are many more options available to the user, other than increased storage capacity.

22.9 Device and Application Constraints

While the AAC device shows much promise, there are still areas where it could be more robust and expanded upon going forward.

22.9.1 Computational Limitations

The Raspberry Pi's inherent computational limitations do not make it feasible to run a powerful text to speech synthesis locally on the device. The text to speech engine used on the Linux version of the AAC device, `Espeak`, is somewhat robotic sounding and does not have many options available to it, causing it to have a mean low opinion score, "which is broadly used to measure the naturalness of the generated speech" [5].

Due to the software limitation laid out by the requirements, many of the higher opinion score text to speech engines cannot be used offline, especially on a low-powered device like the Raspberry Pi. Many of the best text to speech engines either need to run powerful local software or use an online API that allows developers to have the speech synthesis computed in the cloud by a provider.

Due to the increased computational power, both of these techniques can have speech synthesis with very high opinion scores. Some examples of prime online text to speech APIs are Google's Text-To-Speech, IBM's Watson Text to Speech, Mozilla TTS, Microsoft Cognitive Services, Amazon Polly, and more. Other versions of the AAC device can be configured to use these online Text-To-Speech APIs, but that would require the project's scope to broaden outside of the requirements.

22.9.2 Hardware Limitations

The AAC device's following hardware limitations are based on comparing to the device it is trying to compete with, the iPad even though it is in a much lower price tier. The Raspberry Pi AAC device and housing are much thicker than an iPad and can be more challenging to carry around. This thickness is due to a few reasons. The first is that Apple can afford the expense of custom hardware depending on the type of device they are manufacturing. Their net profit as of September 2019 was \$55.256 billion [6]. This AAC device was made with purely off the shelf parts, and these parts were made with low cost in mind. Usually, to make something smaller and thinner, there is an expense to that design decision.

These limitations could be mitigated by using a Raspberry Pi zero and a smaller screen assembly, at the expense of power and ease of use. The touchscreen is another hardware limitation of the device. Apple's smallest iPad, the iPad Mini offers a 7.9 "touchscreen with a resolution of 2048x1536 pixels. The Raspberry Pi AAC device has a 7" touchscreen, with a resolution of 800x480 pixels. This is a considerable decrease in screen resolution and may make it difficult for some users to read small text. This, however, was taken into consideration when designing the GUI of the AAC application, and all text is scaled to prevent eye strain during use.

22.10 Conclusion and Future Plans

There are aspects of the Software and Hardware that can be changed to address some of the limitations discussed in Section 7. To address the thickness of the device, instead of using a smaller Raspberry Pi and screen, a custom 3D printed case could be created to cut down on thickness and more efficiently place the components. This change could also decrease the device's cost, as the off the shelf case costs \$27.

These extra cost savings could also address another hardware limitation, the touch screen. There are higher resolution touch screens available for the Raspberry Pi that are more

expensive than the current one that is used with the AAC device.

In the next iteration of the device, the software can be redesigned several ways. The GUI can be fully configured to modify the word dynamically while the device is running. In addition to this, pictures can be added for users who cannot effectively interpret the word lists, or simply those who would feel pictures to be a more accessible medium for communication. The GUI could also be more pleasing to the eye; that way, users would be more drawn to using the device than any other alternatives that might be flashier or easier to use.

The AAC software also could be packaged better on the device. The application and all of its dependencies need to be installed before the user can use the application. The dependencies are also a part of the core operating system, meaning that potential updates could break the application. To address this, the AAC application could be run inside of a Linux container, such as Docker. Using docker would allow for the application to run on any Raspberry Pi device that has Docker installed. The image would come with all of the required dependencies baked into it. This way, updates to the application could be made seamlessly to the device, without the risk of breaking some aspect of the application.

The benefit of using Docker is that there is minimal overhead compared to simply running the binary by itself. That way, the Raspberry Pi would not be starved of resources for the sake of convenience. Docker allows for both convenience and performance.

Going forward, the goal is that eventually, this device can be mass-produced and sold cheaply to those in need. This could assist people in underprivileged areas who cannot afford other AAC device alternatives and/or internet access. This device could also assist people living in third world countries, who usually would not be able to purchase an AAC device. The Raspberry Pi unit could be assembled, the software loaded, and then shipped to these places for plug and play functionality. This is not limited to just individuals, as school districts or other entities could benefit from this AAC device and could buy multiple of them for the price of a single iPad. Overall, the introduction of low cost, open source computing devices makes the future of alternative medical devices an exciting one.

References

1. Apple Inc, App Store Improvements. (2020). Retrieved September 02, 2020, from <https://developer.apple.com/support/app-store-improvements>
2. Augmentative and Alternative Communication, Retrieved September 02, 2020, from <https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942773> (2020)
3. A. Bondy, L. Frost, The picture exchange communication system. *Behav. Modif.* **25**, 725–744 (2001)

4. A.S. Bondy, L.A. Frost, *A picture's Worth: PECS and Other Visual Communication Strategies in Autism* (Woodbine House, Bethesda, 2002)
5. T.D. Chung, M. Drieberg, M.F.B. Hassan, A. Khalyasmaa, End-to-end Conversion Speed Analysis of an FPT. AI-based Text-to-Speech Application. In *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*, pp. 136–139. IEEE (2020, March)
6. B. Li, *Evaluation of Financial Risk in Apple Company* (Technical University of Ostrava, 2020)
7. M.D. Hossain, H.U. Ahmed, M.J. Uddin, W.A. Chowdhury, M.S. Iqbal, R.I. Kabir, et al., Autism Spectrum disorders (ASD) in South Asia: a systematic review. *BMC Psychiatry* **17**(1), 1–7 (2017)
8. Y.S. Kim, B.L. Leventhal, Y.J. Koh, et al., Prevalence of autism spectrum disorders in a total population sample. *Am. J. Psychiatr.* **168**(9), 904–912 (2011)
9. L. Liberator, *Low-Tech & High-Tech AAC – AAC – Education – Support: Liberator Pty Ltd.* Retrieved October 29, 2020, from <https://liberator.net.au/support/education/aac/low-tech-vs-high-tech> (2018)
10. C. Ma, T. Wang, L. Shen, D. Liang, S. Chen, D. You, Communication-based attacks detection in android applications. *Tsinghua Sci. Technol.* **24**(5), 596–614 (2019)
11. T. Marmarelli, M. Ringle, The reed college iPad study (2011)
12. J. Murphy, I. Markova, S. Collins, E. Moodie, AAC systems*: obstacles to effective use. *Eur. J. Disord. Commun.* **31**, 31–44 (1996). <https://doi.org/10.3109/13682829609033150>
13. NIDCD, Assistive Devices for People with Hearing, Voice, Speech, or Language Disorders. Retrieved 2020 (2011, December)
14. J. Sigafoos, E. Drasgow, Conditional use of aided and unaided AAC: a review and clinical case demonstration. *Focus Autism Other Dev. Disabl.* **16**(3), 152–161 (2001)
15. M. Wedyan, A. Al-Jumaily, O. Dorgham, The use of augmented reality in the diagnosis and treatment of autistic children: a review and a new system. *Multimed. Tools Appl.* **79**(25–26), 18245–18291 (2020). <https://doi.org/10.1007/s11042-020-08647-6>
16. G. Xu, L. Strathearn, B. Liu, W. Bao, Prevalence of autism spectrum disorder among US children and adolescents, 2014–2016. *JAMA* **319**(1), 81–82 (2018)
17. M. Yeo, L. Jiang, E. Tham, W. Xiong, Evaluation of a low-cost alternative communication device with brain control. *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 229–232. <https://doi.org/10.1109/ICIEA.2015.7334116> (2015)

Flavielle Blanco Marques, Gabriel Fernandes Leal, Giovani Nicolas Bettoni, and Osmar Norberto de Souza

Abstract

The incorporation of bioinformatics data in clinical practice is almost happening, mainly due to the rapid growth and access to next-generation sequencing. Thus, personalized precision medicine receives private and government incentives to improve the quality of life and the well-being of the population in the long term. We explore initiatives that seek to map the genome and integrate them with population clinical data, as well as its benefits. Then, we describe the need to develop advanced protocols for exchanging information and platforms capable of processing clinical and genomic data, seeking to increase the well-being of the patients.

Keywords

Bioinformatics data · Bioinformatics projects · Clinical bioinformatics · Clinical health care · Genomics · Grey literature · Health infrastructure · Next-generation sequencing · Populational data · Public health surveillance

23.1 Introduction

Bioinformatics is an interdisciplinary field that uses computational approaches to create solutions to biological problems [1]. In the last years, the number of studies using bioinformatics data have increased due to next-generation sequencing (NGS). Thus, discussions about the use of bioinformatics in

practice clinics are becoming more common. The application of bioinformatics and the integration of data provided by different data “Omics” with the information registered in electronic health records (EHR) can support clinical decisions [2]. This approach has the aim of better clinical decisions and is entitled personalized precision medicine (PPM) [3].

However, there are several challenges to implementation and integration. One big challenge is related to massive and complex data management and data security. This way, bioinformatics is seeking technologies and computational infrastructure to become possible data manipulation and data analyses increasing PPM [4].

Additionally, PPM is associated with the paradigm change of evidence-based medicine (EBM). EBM allows health professionals to develop strong skills in practice based on scientific evidence [5]. In spite, it conducts population studies that do not consider genetic variability among patients [3].

Thus, the combination of information in PPM might better the way patients are classified. Moreover, the diagnosis and treatments will become more efficient [2]. For this to happen, initiatives around the world seek to collect from several data sources information about patients. It happens to have little complete data because missing data in EHR still is a reality. It will allow new ways of monitoring the profile of diseases or their progression [6]. For instance, in the case of epidemics like COVID19. Hence the availability of health data opens opportunities for PPM and bioinformatics, becoming the key to diagnosis and treatments in clinical practice [4].

Therefore in this study, we discuss the advances in personalized precision medicine, public and private initiatives, fields of bioinformatics with direct application in clinical routine or the surveillance of epidemics and public health, patient benefits, and challenges that bioinformatics, PPM, and information technology need to confront nowadays.

In this paper, we highlight relevant concepts and background information (Sect. 23.1), describe the methodology and kind of sources used in our review (Sect. 23.2), delineate

F. B. Marques (✉) · G. F. Leal · G. N. Bettoni · O. N. de Souza
Pontifical Catholic University of Rio Grande do Sul, School of
Technology, Porto Alegre, Brazil
e-mail: flavielle.marques@edu.pucrs.br; gabriel.leal98@edu.pucrs.br;
giovani.bettoni@edu.pucrs.br; osmar.norberto@pucrs.br

main initiatives across the globe and challenges and perspectives (Sect. 23.3). Finally, we discuss our conclusions and final considerations (Sect. 23.4).

23.2 Materials and Methods

Due to the recent development of bioinformatics applications to use in clinics, there are several difficulties to find information about the field. Also, a part of projects come from private initiatives to apply in industry. As these characteristics are on the research theme, in this study, we used the Grey Literature paradigm to carry our review [7]. In this way, it is possible to identify emerging research topics and their applications. Additionally, the information in different sources, outside of academic publishing, can be explored and accepted.

Grey Literature is increasing mainly in the field of software engineering and hopes to fill the gap between academic research and people. Thus, it is possible to include different research sources such as pre-prints, technical reports, and web sites. It allows exploration and understanding like bioinformatics are applying in the clinical context.

Hence, in this study, we realized a seek for materials that describe clinical bioinformatics, concepts, and projects or applications in the field. Additionally, we explored the evolution of bioinformatics and how we can use it. Among the studies, we selected the paper “How can bioinformatics contribute to the routine application of personalized precision medicine?” [3] as the start point of our exploration. The authors established some main terms related to this field and describe an overview of applications of bioinformatics as well as requirements and challenges associated that supported us.

Additionally, the paper “Precision medicine needs pioneering clinical bioinformaticians” [4] was selected as well. It shows a set of initiatives that are using clinical bioinformatics on a national scale with international partnerships. These

initiatives were investigated to understand how bioinformatics is used in each of them.

23.3 Results and Discussion

Clinical bioinformatics development is newer if compared to other health areas. Nevertheless, the field has a lot of future possibilities because it has a big gap for innovation in how it monitors, treats, and stores biological data linked to diseases. Through PPM, the diagnosis and therapeutic can be more effective in the long term, the costs will be reduced for private companies and governments. Consequently, the final costs to patients will be reduced either.

However, the missing data in health records makes communications between clinic and bioinformatics difficult. Therefore, several initiatives around the world have been implemented to fill this data. They are described in Table 23.1. Among them, we can highlight those that are seeking to map genomes and integrate them with clinical data of populations in countries and continents like Asia, the Americas, and the European Union. For this to happen, it is essential encourage data sharing among institutions and countries.

Additionally, 21 countries signed declarations to share data from at least one million genomes by 2022 [8]. The European 1+ Million Genomes Initiative wants to use this information to medicament development, therapies, and interventions more personalized. Therefore, prevention actions, diagnosis, and cost reduction are aims to get. Consequently, benefits in efficiency, accessibility, and health system sustainability are expected.

In 2013, the 100,000 Genomes project was initiated in England, being one of the first in large-scale whole-genome sequencing (WGS) along with clinical data. For this project, were collected DNA samples and clinical data. Besides, the participants received their results and analysis between 2018 and 2019.

Table 23.1 Main initiatives around the world were created to enable the implementation and communication among clinical data, bioinformatics data, and EHR (authors)

Initiative/project	Origin	Year	URL
All of Us Research Program	EUA	2015	https://www.researchallofus.org/
GenomeAsia 100k	Asia	2016	https://genomeasia100k.org/
100,000 genomes project	England	2013	https://www.genomicsengland.co.uk/
ZonMw project	Netherlands	2017	https://www.wgs-first.nl/en
Swiss Pathogen Surveillance Platform (SPSP)	Switzerland	2018	https://www.spsp.ch/
European 1 million genomes initiative	European Union	2018	https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative
ICGC Argo	North America, Europe, Asia, and the Middle East	–	https://www.icgc-argo.org/
Nextstrain	Global collaboration	2018	https://nextstrain.org/

In Switzerland, a national program was developed by the Swiss Institute of Bioinformatics (SIB). It is using bioinformatics data for research, epidemic surveillance, and diagnosis [9]. The initiative has two main aims, the first is to offer resources for health professionals to establish common clinical interpretations of cancer variants identified in patients. The second is to collaborate with hospitals and veterinary institutions in the development of the Swiss Pathogen Surveillance Platform (SPSP) to achieve the complete genome sequencing of bacteria and other multi-resistant pathogens. SPSP allows sharing detailed data for surveillance of outbreaks of pathogens in real-time.

In the Netherlands, the NGS advances have generated incentives in research projects. One field that has been increasing is rare genetic diseases [10]. The ZonMw project aims to assess WGS's usefulness as the first approach to clinical diagnosis includes one evaluation about performance, cost-effectiveness, and impact on medical decisions. As the WGS produces a large amount of information for each individual, a broad and detailed assessment is needed to guide and to judge the practical application of WGS in health care and future personalized treatments [11]. Currently, the main challenge to introduce WGS in rare disease diagnosis is not the technology itself but to integration among clinical knowledge and relevant information from stakeholders (e.g., patients, families, health professionals, and healthcare providers).

In 2018, RADICON-NL (a Dutch consortium) built an X-omics research infrastructure around the country with the capacity to support researchers in data analyses, integration, and management. Additionally, there are projects such as Nextstrain, which has a collaboration of the scientific community to help decision-making for public health [6]. Nextstrain is an open-source project which uses genomics data to get an overview of different diseases, seeking to understand epidemiological data to improve responses to outbreaks for diseases like tuberculosis, dengue, zika, and most recently, Covid-19. All of them, with public data, can be analyzed and visualized to show us the evolution of pathogens and the spread of epidemics.

In the United States, the "All of Us" initiative was designed to increase the number of genome sequencing in the population in a way that this information could be linked to clinical data to understand and determine how a disease can be prevented or treated ("The 'All of Us' Research Program", 2019). The project promotes data sharing and access to patients, doctors, and researchers. Their goal is to reach over 1 million participants in the United States. As of July 20, 2020, the program had about 351,000 participants, more than 220,000 electronic health records, and more than 277,000 bio-samples. Also, the records include both healthy participants and those suffering from a disease, including people who have historically been underrepresented in clinical

research. The diversity of the participants contributes to a wide variety in the types of data, bringing the collected data closer to reality.

GenomeAsia 100k presents itself as a non-profit consortium with the objective of sequencing 100,000 genomes of Asian individuals in an attempt to accelerate the advances in precision medicine, especially for the Asian population [12]. The project foresees the study and availability of data collected from 64 Asian countries in an attempt to improve the diversity of populations in genetic studies and, thus, to investigate variations between genes responsible for diseases in different ethnicities.

Furthermore, there is the ICGC Argo project, which operates in 12 countries on different continents and seeks to explore bioinformatics data to advance cancer-related studies. The initiative addresses different types of cancer and highlights the use of clinical data from more than 80,000 volunteer patients, intending to reach at least 200,000 patients in the future. The approach focuses on the patient and how to improve his quality of life using this data. The issues that the ICGC Argo seeks to understand through clinical bioinformatics are how to use current treatments, changes in cancer cases over time and treatment, how to implement new approaches in health and drug development sectors, and detection and prevention of cancer [13].

23.3.1 Challenges and Perspectives

There are immense challenges to these initiatives, whether public or private, so that personalized precision medicine is widely implemented, especially in the hospital environment. Therefore, the main challenges that we highlight in Table 23.2 are organizational, ethical, and regulatory, on sharing data and infrastructure, in collecting and treatment of genomic information, and economic challenges [4].

Even so, when it is done, each individual should benefit from MPP including improvement in the detection and monitoring of diseases; identification of pre-symptomatic or asymptomatic individuals; modeling the evolution of the disease, among other benefits, as we describe in Table 23.2. Thus, MPP does not have the main objective to increase life expectancy but to improve the quality of life and long-term well-being. Moreover, the possibility of implementation and collaboration around the world contributes to the democratization of health [14].

Moreover, topics such as data security, tools development appropriated for decision making, qualification for medical staff, and higher confidence of patients and medical staff need to be discussed before any solution implementation. Thus, international initiatives are developing guidelines for the acquisition, manipulation, sharing, and use of genomic and clinical data.

Table 23.2 Challenges and opportunities for personalized precision medicine combined with clinical bioinformatics data (authors)

Challenges	Benefits
Large and complex data analysis	Improved disease detection and monitoring
Heterogeneity of data types and formats	Identification of pre-symptomatic or asymptomatic individuals
Standardization of semantics with vocabularies, terminologies, and coding	Modeling and monitoring disease progress
Integration of data from different sources	Effectiveness, accessibility, and sustainability of health systems
Secure data storage and sharing, ensuring privacy	A systemic view of medicine, hospital, and clinical processes
Training and qualification for medical staff, patients, and decision-makers	Patient-centered medicine
Ethical, regulatory, and consent issues	Long-term cost savings

In February of 2017, Health Level Seven (HL7) published a guideline entitled “HL7 Domain Analysis Model: Clinical Genomics”. It presents essential use cases for personalized precision medicine including scenarios more common to genetics testing, cancer, and tumor profile, a late development in early childhood, neonatal testing, and clinical trials for newborns. Also, each use case can include several scenarios and alternatives workflows. The initiative aims to facilitate the development of tools and patterns for interoperability [15] and in Brazil we find studies exploring its application for the public healthcare system and data exchange [16].

Another initiative is the Global Alliance for Genomics and Health, that defines politics and technical patterns to responsible sharing of genomics data [17]. Furthermore, they created GA4GH Connect, a strategic plan of five-years to allow access and sharing of genomics and clinical data of million people until 2022. Tools based on GA4GH pattern will enable that research, health, and trade organizations, as well as people use, analyze, and store data.

The MPP needs advanced technologies and processes that provide ways for acquisition, management, analysis, and support data. Concerning infrastructure, many of the European infrastructures are progressing. ELIXIR is one intergovernmental organization that coordinates, integrating, and sustaining bioinformatics resources. They provide all necessary infrastructure such as databases, software, trainee, and storage [18].

Despite several international initiatives, applications of clinical bioinformatics and their integration into medical practice are few in Brazil. In the last few years, the number of programs in bioinformatics has increased, as well as the demand for qualified professionals for research and industry. However, clinical bioinformatics is mainly used in specific applications and it is realized by private companies in partnership with hospitals. For instance, Genomika institution it was created from Israelita Albert Einstein Hospital. It offers different services to genetic and immunological diseases investing in technology and professionals qualification to realize exams for oncology, allergies, pharmacology, cytogenetics, and analysis of inherited diseases.

Similarly, there is Genera, a laboratory specialized in ancestral exams, pharmacogenomics, and microbiome anal-

ysis. The company invests in research through cooperation with the University of São Paulo (USP). Besides, Brazil has companies such as Neoprospecta, that offers services to biotechnology. Among them are tools for infection control and support for decision-making, microbiome analysis, outbreak investigation, and sequencing of genomes and transcriptomes.

In contrast, in other countries with advanced infrastructure, some information from patients with rare diseases or cancer patients are already being shared. It is expected these patients will have immediate clinical benefits [19]. Cancer is one of the diseases with the most highlights among clinical bioinformatics [20]. However, many drugs used in oncological treatment do not consider the patient’s genetic profile, which has relevant information that can assist in therapeutic management [21].

In France, a center based on academic reference centers named CRefIX will be used to develop processes, tools, and technologies. Nowadays, it is functional and some pilot projects store information on patients with rare diseases, cancer, and diabetes [19]. Furthermore, in Australia, two projects were announced. Diseases more common in the Australian population were prioritized [22]. The first is related to cardiovascular diseases and the second is associated with developing a tracking panel for cystic fibrosis carriers, spinal muscular atrophy (SMA), and fragile X syndrome (FXS) [19]. On top of that, other implementations are expected in different economic aspects. For instance, applications to infectious diseases and common genetic disorders such as sickle cell anemia and thalassemia, hypertension, dyslipidemia, stroke, and kidney disease [19].

Accordingly, MPP and all initiatives need bioinformatics, quality of patient data, and preventive and diagnostic approaches. The development of infrastructure, methodologies, and tools to merge genomics and clinical data is necessary and urgent. Therefore, applications need to be capable of processing and to integrate data from several sources. For this, it is fundamental to define patterns and protocols for data exchange that ensure privacy, security and develop friendly tools that can search, visualize and interpret data. Thus, the quality of life of individuals will be better and also prevention, diagnosis, and treatments will be improved [4].

23.4 Conclusion

The recent advances in sequencing technologies have enabled the growth of clinical applications of bioinformatics, creating new opportunities for innovation and improvement in health processes. In several places, new initiatives appear that aim to incorporate these studies and the information that bioinformatics data can generate. As pointed out by the covered work, are still found challenges in developing this type of solution, involving information security, the amount and datatype of raw, and information sources. However, the benefits that applications related to clinical bioinformatics can bring are as diverse as improvements in treatments, access to health resources, the development of new pharmaceuticals, as well as economic benefits to businesses and governments.

Currently, initiatives are being implemented globally to use different ways of bioinformatics data in the clinic. Among the approaches that stand out the most, we highlight those that seek to share information related to genetics, use in epidemiological control and monitoring, genetic mapping and study of different populations, and treatment of inherited diseases and oncology. From the studies explored, we highlighted the need to share information between projects, companies, and research centers. This seems to be a common goal for many initiatives and the best way to increase the growth of this kind of application and discoveries in the area.

Moreover, concerning initiatives in infrastructure development, standards, and protocols, methodologies, and tools capable of joining clinical and genomic data are found at different stages. Given this scenario, these demands need training and experience in handling and studying bioinformatics data. Nevertheless, related to data protection, the qualification of the professionals, the appropriate tools to assist implementation of any decision-making. Thus, a custom precision medicine, based on the initiatives of family medical bioinformatics around the world, can improve the quality of life with the prevention, diagnosis, and targeted treatments raising the long-term well-being of the population.

Acknowledgments The authors would like to also thank CNPq and CAPES for their financial support. Also, they gratefully thank the Pontifical Catholic University of Rio Grande do Sul, and all of which have helped improve this work.

References

1. N.M. Luscombe, D. Greenbaum, M. Gerstein, What is bioinformatics? A proposed definition and overview of the field. *Meth. Inf. Med.* **40**(04), 346–358 (2001)
2. K. Shameer, M.A. Badgeley, R. Miotto, B.S. Glicksberg, J.W. Morgan, J.T. Dudley, Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Brief. Bioinform.* **18**(1), 105–124 (2017)
3. C. Carretero-Puche, S. García-Martín, R. García-Carbonero, G. Gómez-López, F. Al-Shahrour, How can bioinformatics contribute to the routine application of personalized precision medicine? *Exp. Rev. Prec. Med. Drug Development* **5**, 115–117 (2020)
4. G. Gómez-López, J. Dopazo, J.C. Cigudosa, A. Valencia, F. Al-Shahrour, Precision medicine needs pioneering clinical bioinformaticians. *Brief. Bioinform.* **20**(3), 752–766 (2019)
5. R.P. El Dib, Como praticar a medicina baseada em evidências. *J. Vasc. Brasil.* **6**(1), 1–4 (2007)
6. J. Hadfield, C. Megill, S.M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R.A. Neher, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**(23), 4121–4123 (2018)
7. V. Garousi, M. Felderer, M.V. Mäntylä, Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol.* **106**, 101–121 (2019)
8. G. Saunders, M. Baudis, R. Becker, S. Beltran, C. Bérout, E. Birney, C. Brooksbank, S. Brunak, M. Van den Bulcke, R. Drysdale et al., Leveraging european infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* **20**(11), 693–701 (2019)
9. A. Egli, D.S. Blanc, G. Greub, P.M. Keller, V. Lazarevic, A. Lebrand, S. Leib, R.A. Neher, V. Perreten, A. Ramette et al., Improving the quality and workflow of bacterial genome sequencing and analysis: paving the way for a Switzerland-wide molecular epidemiological surveillance platform, *Swiss Med. Week.* **148**, w14693 (2018)
10. J. Bremer, E.H. Van der Heijden, D.S. Eichhorn, R. Meijer, H.H. Lemmink, H. Scheffer, R.J. Sinke, M.F. Jonkman, A.M. Pasmooij, P.C. Van den Akker, Natural exon skipping sets the stage for exon skipping as therapy for dystrophic epidermolysis bullosa. *Mol. Therapy-Nucleic Acids* **18**, 465–475 (2019)
11. G. Matthijs, E. Souche, M. Alders, A. Corveleyn, S. Eck, I. Feenstra, V. Race, E. Sistermans, M. Sturm, M. Weiss et al., Guidelines for diagnostic next-generation sequencing. *Eur. J. Hum. Genet.* **24**(1), 2–5 (2016)
12. G. Consortium et al., The genomeasia 100k project enables genetic discoveries across asia. *Nature* **576**(7785), 106 (2019)
13. ICGC ARGO, Jul 2020 [Online]. Available: <https://www.icgc-argo.org/>
14. J.S. Beckmann, D. Lew, Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome Med.* **8**(1), 134 (2016)
15. R. Dolin, A. Boxwala, J. Shalaby, A pharmacogenomics clinical decision support service based on FHIR and CDS hooks. *Methods Inf. Med.* **57**(S02), e115–e123 (2018)
16. C.A.C. Bezerra, A.M.C. de Araújo, V.C. Times, An hl7-based middleware for exchanging data and enabling interoperability in healthcare applications, in *17th International Conference on Information Technology–New Generations (ITNG 2020)* (Springer, New York, 2020), pp. 461–467
17. Global Alliance for Genomics and Health, A federated ecosystem for sharing genomic, clinical data. *Science* **352**(6291), 1278–1280 (2016)
18. R. Drysdale, C.E. Cook, R. Petryszak, V. Baillie-Gerritsen, M. Barlow, E. Gasteiger, F. Gruhl, J. Haas, J. Lanfear, R. Lopez et al., The elixir core data resources: fundamental infrastructure for the life sciences. *Bioinformatics* **36**(8), 2636 (2020)
19. Z. Stark, L. Dolman, T. A. Manolio, B. Ozenberger, S.L. Hill, M.J. Caulfield, Y. Levy, D. Glazer, J. Wilson, M. Lawler et al., Integrating genomics into healthcare: a global responsibility. *Am. J. Hum. Genet.* **104**(1), 13–20 (2019)

20. B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz, K.W. Kinzler, Cancer genome landscapes. *Science* **339**(6127), 1546–1558 (2013)
21. C. Eifert, R.S. Powers, From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets. *Nat. Rev. Cancer* **12**(8), 572–578 (2012)
22. A.D. Archibald, M.J. Smith, T. Burgess, K.L. Scarff, J. Elliott, C.E. Hunt, C. Barns-Jenkins, C. Holt, K. Sandoval, V.S. Kumar et al., Reproductive genetic carrier screening for cystic fibrosis, fragile X syndrome, and spinal muscular atrophy in australia: outcomes of 12,000 tests. *Genet. Med.* **20**(5), 513–523 (2018)

Derek Stratton, William Garner, Terra Williams, and Frederick C. Harris Jr.

Abstract

A pandemic can arise without warning, and it is important for those in charge of managing the outbreak to understand how diseases spread. Being able to simulate the spread of a disease in varying environments can help the world be more prepared when an outbreak occurs. The COVID-19 City Simulator allows the user to test the spread of the virus under multiple different scenarios. Parallel computing can help to make these simulations more efficient by allowing data to be gathered at a faster rate on a particle simulation. This paper shows how OpenMP and MPI can improve a pandemic simulation by cutting the runtime from over 25 s to under 10 s when 4 threads and 4 boxes are used. We also find that the speed of implementing a lockdown largely impacts the amount of cases and deaths in the city.

Keywords

Parallel computing · OpenMP · MPI · Hybrid model · COVID-19 · Pandemic · Simulation · disease spread · Particle simulation · Parameter tuning

24.1 Introduction

Many mathematical models exist to represent the spread of infectious diseases in a population. These models vary widely, but all models have large limitations when gener-

alizing a population. However, they can still be helpful to gain understanding of the biological and sociological factors that contribute to the spread of disease [1]. This knowledge can be used to advice public health policy so institutions can deal with disease outbreaks in the best way possible. This study is being conducted due to its modern relevance with the COVID-19 pandemic.

Many countries have reacted differently to this pandemic, but being able to accurately simulate the spread of a disease can help the entire world handle outbreaks more effectively. This simulation can display how different factors such as population density and amount of public interaction will affect the spread of a virus. We will be looking at lockdowns orders, which include public policies that attempt to reduce spread by reducing interactions in a city. Due to the challenge of implementing lockdowns, some countries have implemented them more strictly and more quickly than others. This study attempts to investigate the extent to which the speed at which a city locks down affects the public health outcomes of the city.

The paper will also discuss how parallel computing can make simulations more efficient, since making simulations more efficient will allow for them to be more realistic and useful for understanding the problem. Shared memory parallelism involves using multiple threads that share a common memory space to execute tasks in parallel. Message passing parallelism involves executing tasks on different boxes that don't share memory, and require messages be sent to exchange information needed for processing. Both techniques are powerful and can be used in tandem to maximize the efficiency of computing resources on a simulation.

In Sect. 24.2, the background and relevant past research are discussed. In Sect. 24.3, the detailed approach of the simulation's implementation is described. In Sect. 24.4, the results of simulation trials are reported and analyzed. Finally, Sect. 24.5 gives the conclusion of the paper's work and describes future work to expand knowledge on this topic.

D. Stratton · W. Garner · T. Williams · F. C. Harris Jr. (✉)
 Department of Computer Science and Engineering, University of Nevada Reno, Reno, NV, USA
 e-mail: derekstratton@nevada.unr.edu;
williamgarner@nevada.unr.edu; terrawilliams@nevada.unr.edu;
fred.harris@cse.unr.edu

24.2 Related Works

The simulation discussed in this paper is based on other models used to simulate the spread of disease. The first model used as inspiration is the SIR disease model. This model splits a population into three compartments: susceptible, infected, and recovered. This model was based on Kermack-McKendrick theory from 1927 [2]. The SIR Disease model uses transition functions to move individuals of the population between the three compartments. This system of compartments and transitions is modeled with ordinary differential equations (ODEs), and it is a deterministic model. Another type of models that improve upon deterministic models are stochastic disease models. They improve upon standard compartmental models like SIR by adding some elements of randomness [3]. These stochastic models can use Continuous Time Markov Chains or Stochastic Differential Equations to simulate the spread of disease.

Britton describes a special epidemic model with site contamination [4]. This model divides a space into several different sites with a random number of particles in each site to represent the individuals of a population. The particles can move randomly between neighboring sites. When an infected particle reaches a new site, all other particles currently at that site become infected. Germann made a model that used mitigation strategies for pandemic influenza in the United States [5]. This was a complex simulation model used to study influenza in the United States. The model used data from the 2000 US Census and divided the population into seven “mixing groups” that could contact one another. The model considered various interventions such as vaccination and social distancing to determine how different factors would affect the spread of influenza.

For the simulation, multiple methods of computer parallelization of particle simulators were researched. We had to decide which method of parallelization would be most appropriate for the project. One source of research discussed many different methods and platforms for parallel computing [6]. From this paper as well as our former experiences, we decided that MPI was the most appropriate method for parallelizing the simulation. Our research into parallelization methods also showed the difference in performance for a program using pure MPI against a program using a hybrid of MPI and OpenMP [7]. The hybrid uses of both MPI and OpenMP was found to improve performance over using MPI alone. The combination of these two sources lead us to test the simulation using three different methods of parallelization: OpenMP, MPI, and a hybrid model using both OpenMP and MPI.

24.3 Approach

The city is represented by a particle simulation. The particles exist in a finite 2D space where particles represent people living in the city and the space represents the city. The simulation moves forward at discrete time steps, where each time step t represents transitioning an hour forward in real time.

To create a simulated epidemic model of the city, various properties are added to both the particles and the organization of the city to help simulate how a real city would experience an epidemic.

24.3.1 Particles with Disease States

Each particle in the simulation will always exist in a disease state. The set of states, \mathcal{S} , includes Susceptible (S), Infected (I), Recovered (R), and Deceased (D), based off the SIR compartmental model (24.1).

$$\mathcal{S} \in \{S, I, R, D\} \quad (24.1)$$

We let N_t represent the number of total particles in the simulation at a given time step. All living individuals will be represented as particles in the simulation (24.2). Since deceased individuals will be removed from the simulation, N can decrease over time.

$$N_t = |S_t| + |I_t| + |R_t| - |D_t| \quad (24.2)$$

The number of initial particles in the simulation, N_0 , is a variable that will be examined as a parameter to determine how larger population size (and thus higher population density) affects the disease transmission statistics. The simulation begins with all individuals except for 1 being susceptible (24.3), 1 infected individual, and no recovered or deceased individuals.

$$|S_0| = N_0 - 1 \quad (24.3)$$

Susceptible individuals have the chance to transition to the infected state at every time step. Infected individuals can spread the disease to susceptible individuals. The chance of infection from an infected individual to a susceptible one at any given time step is given by the probability of infection α . However, infectious particles can only affect susceptible particles if they are within a certain radius of infection β (24.4). Individuals must also be in the same area to infect each other, which will be discussed more in section B on City Organization.

$$P(x \text{ infecting } y) = \begin{cases} \alpha & \text{dist}(x, y) \leq \beta \\ 0 & \text{otherwise} \end{cases} \quad (24.4)$$

Infected individuals can transition to either the recovered or deceased state with probabilistic functions based on a recovery factor γ and death factor δ , respectively (24.5), (24.6). These functions are also based on the time since infection, with a higher chance of either recovering or perishing the longer they are infected. These probabilities are tested at each time step.

$$P(\text{recovery, } i \text{ steps after infection}) = \max(\gamma i, 1) \quad (24.5)$$

$$P(\text{death, } i \text{ steps after infection}) = \max(\delta i, 1) \quad (24.6)$$

Recovered individuals are still part of the population but cannot transition out of the recovered state. Deceased individuals are fully removed from the population and simulation. The simulation ends either when time step t reaches some stopping point T or when there are no infected individuals remaining in the population, whichever happens first.

24.3.2 City Organization

The total space of the city is partitioned into units we call areas. Areas represent discrete regions where people stay, and they are each represented as equal size, square regions. There are both public and personal areas. Personal areas represent homes, and every person will have exactly one personal area that they visit. Public areas represent places that many people visit, such as schools, workplaces, or stores.

The number of areas is proportional to the number of particles in the simulation. For every 2 initial particles in the simulation there will be 1 personal area, which is a proportion that reflects housing to population data for United States cities. For every 4 personal areas in the city, there will be 1 public area. Figure 24.1 shows an example of the box distributions. The relative locations of areas do not matter in this simulation.

At the beginning of the simulation, a set of areas are assigned to each person. Personal areas are assigned such that every person belongs to 1 personal area and every personal area has 2 people. Public areas are assigned with more variability, with each being assigned to $N_0/10$ random people. These values are treated as constant controls as other parameters are investigated for their effects.

Each area that a person has assigned to them will also come with a probability p of transitioning there at any given frame.

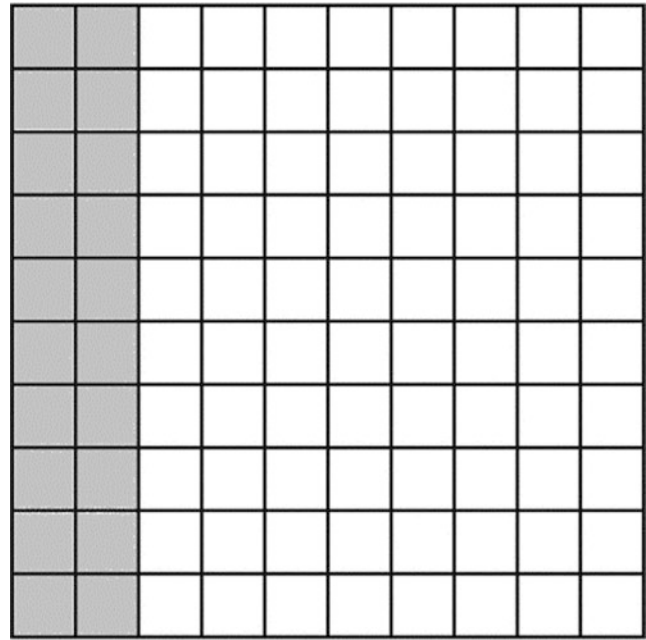


Fig. 24.1 Example 10×10 grid layout for a city with $N_0 = 160$. The shaded regions represent public areas and the unshaded represent personal areas

24.3.3 Simulating Movement of People

There are two types of movements that particles have in the simulation: movement between areas and movement within an area.

Movement between areas is considered an instantaneous jump from one area to another. At each time step, every person will have a chance p to jump to one of their assigned areas (assuming they are not already in that area). A random real value between 0 and 1 is uniformly drawn for every person's possible areas to jump to, and if the value is less than the probability, the jump will occur. The particle's position will be randomly chosen in the new area.

It is worth noting that since these jump timings are random, people do not have set "schedules" and can stay in areas for short and long amounts of time. There is also no notion of day and night. The values are chosen to best simulate people on average dividing their time evenly between time in public and their time at home.

An intervention policy that is tested to affect these probabilities is the lockdown. With a lockdown, the goal is to limit the number of jumps to public areas. This is accomplished by reducing the probability that any particle will jump to a public area. A parameter that will be tested to find the effect of this is the time of intervention implementation, τ . Changing this

Table 24.1 Parameters that will be investigated and tuned in the simulations

Symbol	Description
N_0	The number of initial particles
α	The probability of infection at a time step
β	The radius of infection for 2 particles
γ	The factor for recovery probability
δ	The factor for death probability
p	The probability of jumping to a new area
τ	The time of lockdown intervention

will be used to determine the success of implementing the lockdown early as to opposed to late.

Movement within an area occurs when a particle is not jumping between areas. The movement occurs by randomly picking an x and y value, each with a maximum magnitude. This type of movement is used to simulate people coming in proximity with one another to potentially spread the disease.

24.3.4 Parameter Estimation

To create a useful model, different parameters of the system must be chosen that help to describe phenomena that occur in real life. Table 24.1 shows a list of these parameters that will need to be decided. They will be decided using both logic and with respect to other parameters in the simulation. For example, the probability of infection is to be chosen by using the current knowledge that COVID-19 has a R_0 (Reproduction Number) number of approximately 2.2. A value for α will be decided so that the trend shows that around 2.2 particles are infected by a single infectious particle.

24.3.5 Parallelization Strategy

The simulation is implemented in C++ while utilizing the OpenMP and MPI libraries for both shared memory and message passing parallelization. Testing will be done on the Bridges supercomputing environment where we will access up to 16 nodes to test with at a time.

To distribute work evenly across the nodes used in the simulation, every node will get the same number of public and personal areas. Since the particles can jump to an area on any node, the communication of jumping particles must be done sequentially with each node having a chance to send its jumping particles while all others must be ready to receive.

Since relative position between areas doesn't matter, each area can be viewed as its own 2D space, and the relative (x , y) coordinate of each particle in the area, and the id of the area where the particle exists is sufficient information for processing.

24.3.6 Strategy for Measuring Success

Two approaches are used to measure the success of our simulation. The first is to analyze the infections, recoveries, and deaths that occur in the simulation. The second is to measure the impact of parallelization on the simulation's performance.

For the first approach of analyzing disease metrics, graphs of infections, recoveries, and deaths over time will be produced. Graphs will also be produced that measure the total number of infections, recoveries, and deaths that are caused by different values. This provides useful intervention insight in the case of altering τ , since we can measure how many lives can be potentially saved by how quickly the lockdown is implemented following the disease onset.

The second approach is to analyze the speedup induced by the parallelization of the simulation. Experiments for both strong and weak scaling will be performed by changing the number of total particles in the city to show that the hybrid approach scales well compared to a sequential implementation.

24.4 Results

One of our goals when running the simulation was to determine the effectiveness of parallelization using OpenMP, MPI, and then a hybrid of the two. Having a simulation run efficiently is important for its usefulness as a tool to better understand the spread of disease. First, we tested the application using only OpenMP for parallelization. We ran a simulation of 100,000 people using 1, 2, 4, 8, and 16 threads. We ran each of these simulations 10 times and used the average run time to determine the speedup of the simulation with OpenMP. Figure 24.2 shows the average time taken to run the simulation for each number of threads. Figure 24.3 shows the speedup of the simulation with each number of threads.

Next, we tested the simulation using only MPI for parallelization. Similarly to the OpenMP tests, we tested the MPI simulation with 100,000 people and 1, 2, and 4 processors. Unfortunately, due to resource restrictions with the Bridges supercomputer at the time of these tests, we were not able to run the simulation with more than 4 processors. We ran each of the simulations 10 times and used the average runtime to determine the speedup, strong scaling efficiency, and weak scaling efficiency. Figure 24.4 shows the average runtime of the simulation with different numbers of processors. Figure 24.5 shows the speedup of the simulation with an increasing number of processors.

We were able to see a much larger speedup when using MPI than when using OpenMP. Even with 16 OpenMP threads, the speedup was below 2.5 while 4 processors with

Fig. 24.2 The average runtime of the simulation tends to decrease with additional OpenMP threads

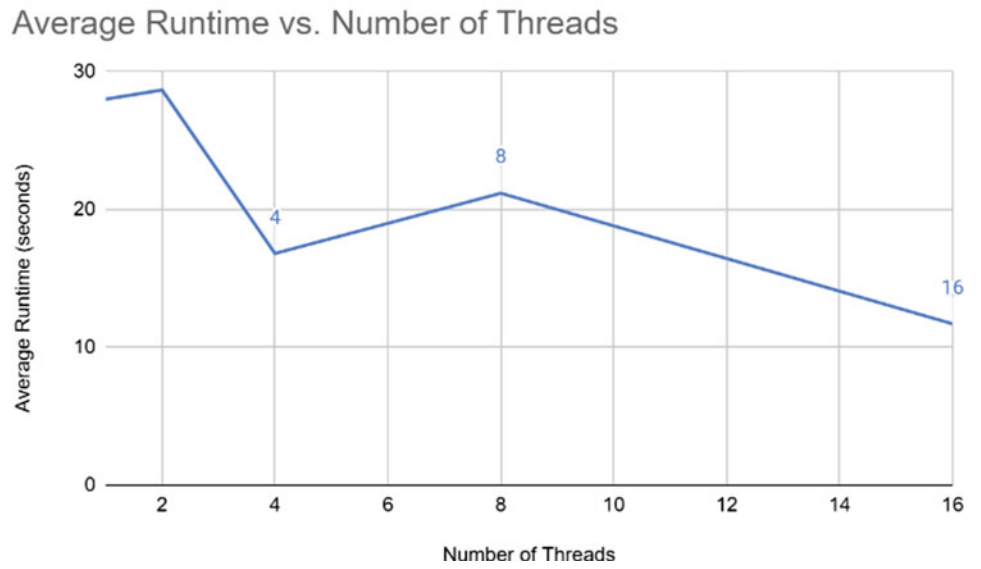


Fig. 24.3 The speedup of the program tends to increase with more OpenMP threads

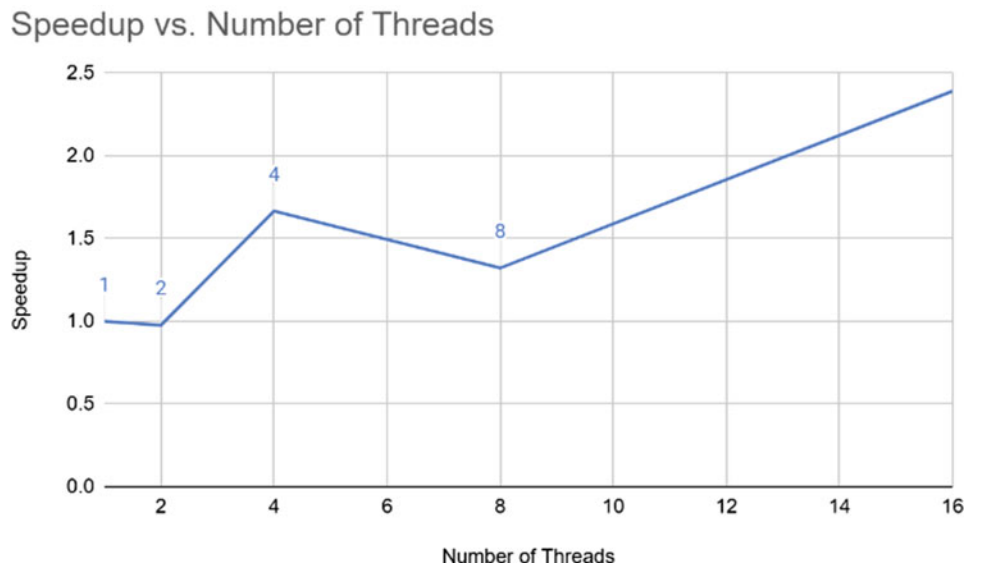


Fig. 24.4 The average runtime of the simulation decreases when more processors are added with MPI

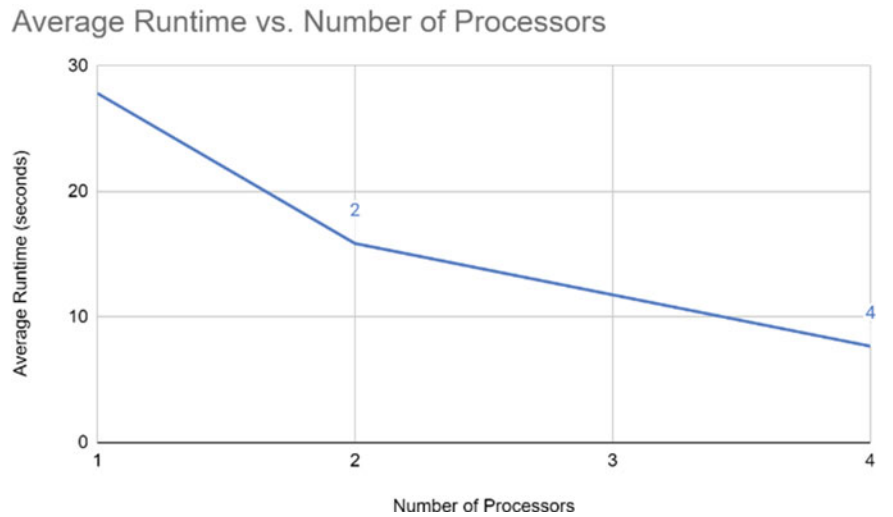


Fig. 24.5 The speedup of the program increases when more processors are added with MPI

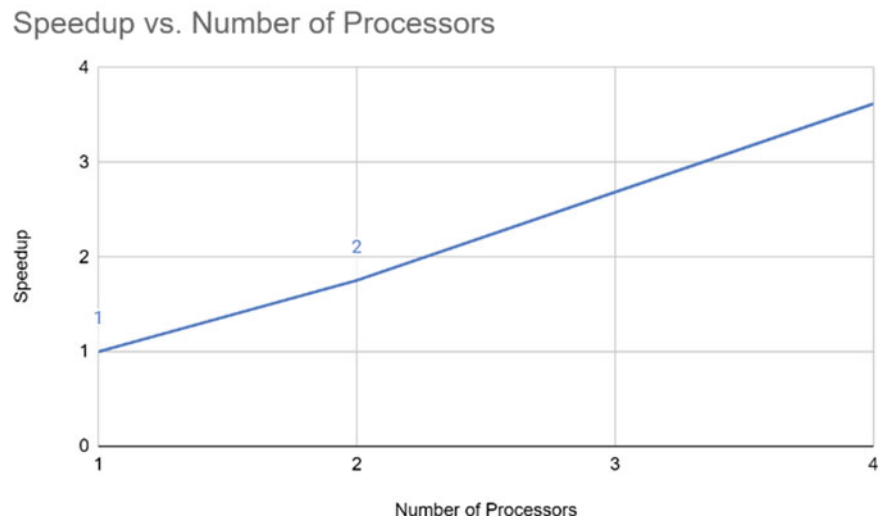
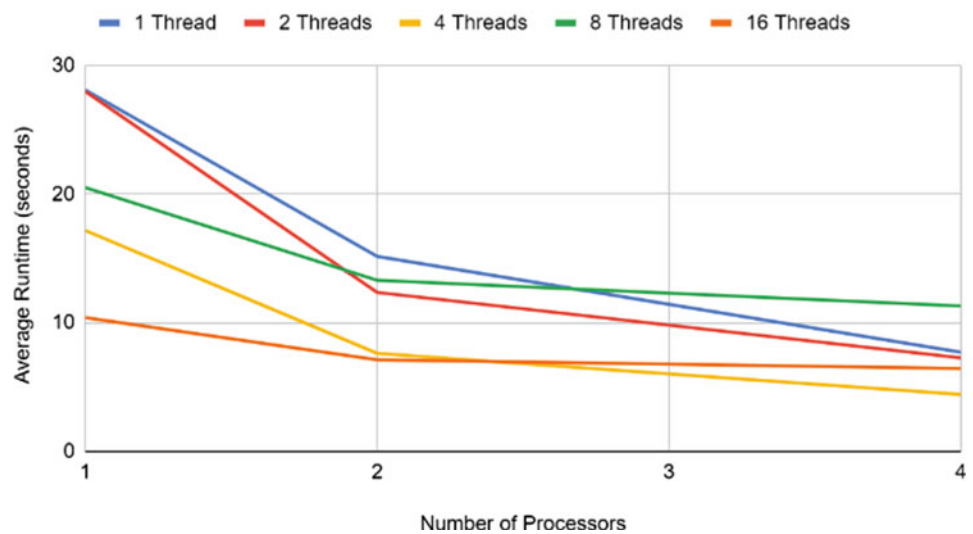


Fig. 24.6 The runtime of the simulation tends to decrease with more threads and more processors



MPI yielded a speedup of about 3.6. The strong scaling efficiency with MPI also stayed above 85% while the strong scaling efficiency of OpenMP dropped to about 42% with 4 threads and to about 15% with 16 threads.

The final test conducted on the efficiency of different parallelization methods was to use a hybrid model of OpenMP and MPI. We ran these tests with 100,000 people in the simulation. We ran the tests with 1, 2, 4, 8, and 16 threads, and 1, 2 and 4 processors. Figure 24.6 shows the resulting runtimes with various configurations of threads and processors. We found the best speedup with 4 threads and 4 processors, but overall, the runtime tended to decrease when more threads and processors were used.

Besides running tests on the efficiency of the simulation, we also ran tests to determine the effectiveness of a lockdown during a disease outbreak. In the simulation, a lockdown can be set to occur during a particular time step. After the lockdown occurs, people will be much less likely to jump

to areas outside of their personal space. Figure 24.7 shows the number of people susceptible, infected, recovered, and deceased for each time tick when no lockdown was implemented. Figure 24.8 shows the same statistics but in this case, a lockdown was implemented about one quarter of the way through the simulation. Figure 24.9 shows the results of implementing a lockdown one tenth of the way through the simulation.

We can see from Figs. 24.7, 24.8, and 24.9 how a lockdown slows the spread of disease. With no lockdown, about 91% of the simulation population became infected at some time. Even when a lockdown occurred a quarter of the way through the simulation, about 91% of the population was infected with the disease. We saw the most dramatic difference when a lockdown was implemented one tenth of the way through the simulation. In this instance, only 20% of the population became infected with the disease.

Fig. 24.7 State of the disease over time when no lockdown is implemented

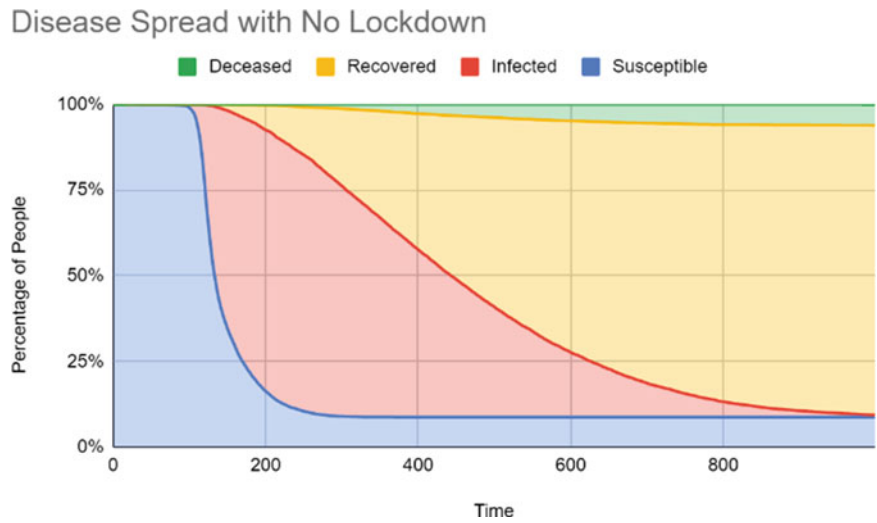


Fig. 24.8 Spread of disease when a lockdown is implemented one quarter of the way into the simulation

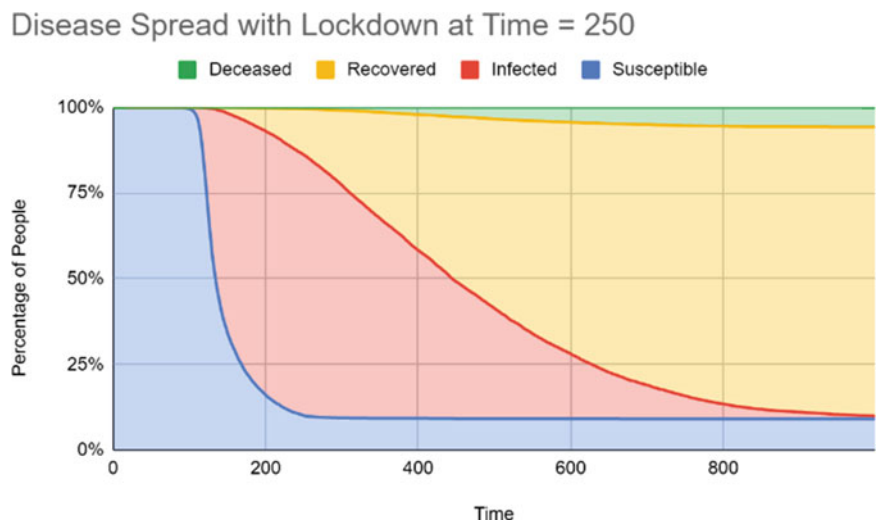
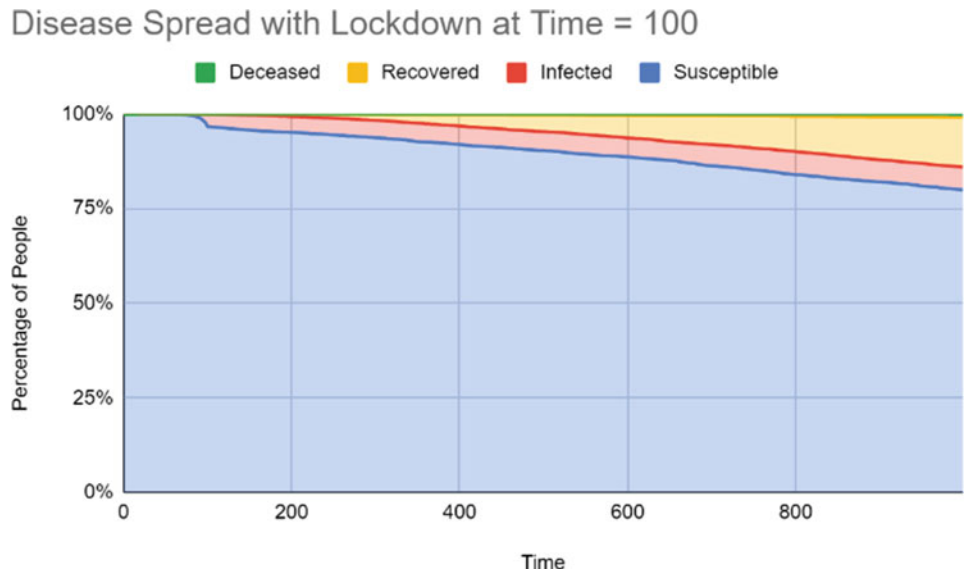


Fig. 24.9 Spread of the disease when a lockdown is implemented early in the simulation



24.5 Conclusions and Future Work

In conclusion, we found that a hybrid parallel model using both OpenMP and MPI yielded the best performance when running the simulation. We also found that using only MPI produced a much better efficiency than using only OpenMP. OpenMP however did improve the runtime over using the base serial implementation.

In terms of improving the performance of the simulation, future work could include implementing different approaches to parallelization. Other technologies such as CUDA could be used to implement the simulation, and then the results could be compared to our implementation. There are still many methods and platforms for parallelization that could be explored with this project [6]. This future work could help to discover if using MPI and OpenMP was the most appropriate way to implement the simulation. If allowed more resources in the future, we would also like to be able to test our simulation using more than 4 processors with MPI. We would like to be able to see how using 8 and 16 processors would improve the speedup of the simulation.

The results of the simulation showed just how important early action is when a possible pandemic is at hand. Early preventative measures proved to be very effective at reducing the number of people to be infected by a disease. Our data showed that waiting too long to try to stop the spread of infection can allow the disease to reach many people.

In the future, more work could be done to make the simulation a better representation of the world's population as our Coronavirus simulation was conducted on a homogeneous population. Possible future work into this simulation could include introducing genetic differences such as age or preexisting health conditions that would make a person more or less likely to die after being infected. Immunities could be introduced so that some people within the simulation are unable to be infected. Areas of different population densities could be introduced to see how the spread of a disease differs in a large community versus a smaller one. Another factor that could be added to the simulation is hospitalization. Infected persons would be isolated to one location therefore

making them less likely to infect others and more likely to recover. Another interesting addition to the simulation would be to add the creation of a vaccine to the disease. This would allow us to see how a vaccine would help to lessen the spread of the disease as well as see how the lifetime of the disease is shortened. There exists a lot of future work that could be done to make this simulation a more accurate representation of the world's population.

Acknowledgments This material is based in part upon work supported by the National Science Foundation under grant number IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. N. Becker, The uses of epidemic models. *Biometrics* **35**(1), 295–305 (1979). ISSN: 0006341X, 15410420 [Online]. Available: <http://www.jstor.org/stable/2529951>
2. W.O. Kermack, A.G. McKendrick, G.T. Walker, A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Char.* **115**(772), 700–721 (1927). <https://doi.org/10.1098/rspa.1927.0118>. eprint: <https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1927.0118> [Online]. Available: <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1927.0118>.
3. L.J. Allen, A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infect. Dis. Modell.* **2**(2), 128–142 (2017). ISSN: 2468-0427. <https://doi.org/10.1016/j.idm.2017.03.001> [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2468042716300495>
4. T. Britton, M. Deijfen, F. Lopes, A spatial epidemic model with site contamination (2017). arXiv: 1705.07448 [math.PR]
5. T.C. Germann, K. Kadau, I.M. Longini, C.A. Macken, Mitigation strategies for pandemic influenza in the United States. *Proc. Natl. Acad. Sci.* **103**(15), 5935–5940 (2006). <https://doi.org/10.1073/pnas.0601266103>. eprint: <https://www.pnas.org/content/103/15/5935.full.pdf> [Online]
6. P. Czarnul, J. Proficz, K. Drypczewski, Survey of methodologies, approaches, and challenges in parallel programming using high-performance computing systems. *Sci. Program.* **2020**, 4176794:1–4176794:19 (2020)
7. M. Hipp, W. Rosenstiel, Parallel hybrid particle simulations using MPI and OpenMP, in *Euro-Par 2004 Parallel Processing*, ed. by M. Danelutto, M. Vanneschi, D. Laforenza (Springer, Berlin, Heidelberg, 2004), pp. 189–197. ISBN: 978-3-540-27866-5

Part V

Management and Applications

Techniques and Tools for Selection and Strategic Alignment of Projects and Projects Portfolio Balancing: A Systematic Mapping

Djenane C. S. dos Santos, Adler D. de Souza, and Flávio B. S. Mota

Abstract

The optimization of resources used in projects, whether in the context of private companies or in the context of public management, is directly related to efficient the selection and strategic alignment of projects and portfolio balancing through appropriate techniques and tools. This work describes a systematic mapping carried out with the purpose of identifying which tools and techniques are used (or are more appropriate) for selection and strategic alignment of projects and project portfolio balancing. The research was conducted from the digital libraries Scopus and IEEE, resulting, initially, in a total of 128 articles. After applying the filters and the exclusion and inclusion criteria adopted, the study was restricted to 11 articles in Scopus and 11 in IEEE, in addition to 1 more article that was included by snowball. The research made it possible to verify that the vast majority of the techniques used are Multi-Criteria Decision Support Methods and models that use fuzzy logic, in addition to Evolutionary Algorithms as the main tool.

Keywords

Strategic alignment · Portfolio balancing · Tools · Techniques · Project selection

D. C. S. dos Santos (✉) · A. D. de Souza · F. B. S. Mota
Postgraduate Program in Science Computer and Technology, Institute of Mathematics and Computation, Itajubá, Brazil

Department of Mathematics and Computation, Federal University of Itajubá, Itajubá, Brazil
e-mail: djenane@unifei.edu.br; adlerdiniz@unifei.edu.br; flaviomota@unifei.edu.br

25.1 Introduction

Considering the growing search for the resources optimization, whether in the context of private companies or in the context of public management, necessary in the development of projects, the proper management and balancing of the project portfolio and strategic alignment are essential points to be sought.

According to [1], the large volume of projects and the frequent changes in the scenarios where the companies are inserted bring with them the need to search for faster and higher quality results, with low costs and in shorter times. Thus, project selection and monitoring are essential to ensure that the resulting project portfolio aligns with the organization's strategies and allows for the creation of business values, adaptation to available capacity and resources and increased stakeholder satisfaction [2].

According to [3], project portfolio management is responsible for alignment, for linking the organization's business strategy and the correct selection of projects that will make it possible to achieve the expected results.

The portfolio is a set of programs, projects or operations managed as a group to achieve strategic objectives. The portfolio components are quantifiable, that is, they can be measured, classified and prioritized [4].

A program is defined by [4] as a group of related projects and which, when managed in a coordinated manner, make it possible to obtain strategic benefits and control that would not be available if they were managed otherwise. The project can be seen as a temporary effort made to obtain results, be it a product, service or other result, which is exclusive.

According to [5], the dynamism implicit in most organizations continually raises the number of uncertainties and risks inherent in their activities, which increases the complexity of the decision-making process related to project selection and prioritization and project portfolio management.

There are several variables involved and that need to be considered when evaluating projects. Changes and uncertainties in the market in which the organization is inserted, resources are almost always scarce, choosing the appropriate criteria for the classification, selection and prioritization of projects are some of the variables that will interfere in the alignment of the portfolio with the organization's strategy. The expected benefits and results may not be achieved if inadequate analysis occurs [3].

In view of the above, the need for systematic mapping is justified in order to characterize which tools and techniques are currently being used for selection and alignment strategic of projects and project portfolio balancing. This work presents in the next sections the systematic mapping process, the description of the stages of systematic mapping, the discussion of research questions and, finally, the conclusion.

25.2 Systematic Mapping Process

Systematic mapping seeks a broader field for any type of research, in order to obtain an overview of the state of the art or practice on a topic [6]. The systematic mapping methodology meets the objective of the article, which is to seek the state of the art in tools and techniques used for the selection and strategic alignment of projects and balancing the project portfolio. The process, adapted from [6], of the systematic mapping of the adopted literature is in the next section.

25.2.1 Stages of Systematic Mapping

25.2.1.1 Objective Definition

Analyze scientific publications through a study based on systematic mapping **in order to** identify which tools and techniques are used (or are more suitable) for the selection and strategic alignment of projects and project portfolio balancing of the researchers **point of view** in **context** of public or private organizations.

25.2.1.2 Research Question Definition

What tools and techniques are used (or are they most appropriate) **selection and strategic alignment of projects and project portfolio balancing?**

25.2.1.3 Defining the Search Strategy and Search Strings

The research was conducted from the digital libraries Scopus and IEEE. The study considered works classified as articles in the period between 2010 and 2020. The following search string was used in the digital library Scopus:

(tools OR techniques) AND (“Project selection” OR “Strategic Alignment” OR “Portfolio Balancing”) AND “project portfolio management”

The following search string were used in the IEEE digital library, with the second search string limiting the search scope to the abstract only:

(((((“All Metadata”: tools OR techniques) AND “All Metadata”: project selection OR strategic alignment OR portfolio balancing) AND “All Metadata”: project portfolio management))

(((((“Abstract”: tools OR techniques) AND “Abstract”: project selection OR strategic alignment OR portfolio balancing) AND “Abstract”: project portfolio management))

25.2.1.4 Definition of the Selection Criteria Inclusion Criteria

For an article to be classified as Accepted, it must meet three inclusion criteria simultaneously, in order to answer the research question. Inclusion criteria are:

- CI-1 The tools and techniques used can be applied in other contexts (or wider contexts).
- CI-2 The results were presented in a satisfactory manner, that is, the data or information allows the formation of a conclusion regarding the tools and techniques used.
- CI-3 The tools and techniques used were applied to a significant sample and/or there are indications that they can be used in another sample.

Exclusion Criteria

For an article to be classified as Rejected it is enough that it meets one of the three exclusion criteria, as this way it will not be possible to answer the research question. Exclusion criteria are:

- CE-1 The tools and techniques used cannot be applied to other contexts (or wider contexts).
- CE-2 The results were not presented satisfactorily, that is, data or information about the techniques/tools used are missing, or in the case of a proposal for a new method/-model/process, there was no application/evaluation of the proposal.
- CE-3 The study did not actually present techniques or tools that have been used for the selection and strategic alignment of projects and project portfolio balancing.

25.2.1.5 Search for Articles and Data Extraction

The first search, still without a period filter and with a search string being applied to the title, abstract and keywords, returned a total of 128 results, 105 in IEEE and 23 in Scopus. After filtering by period, between 2010 and 2020, the search returned a total of 67 results, 52 in IEEE and 15 in Scopus.

Table 25.1 Selected articles in systematic mapping

Refs.	Title	Inclusion Criteria
[7]	The problem of research project portfolio selection in educational organizations: A case study	CI-1, CI-2, CI-3
[8]	Decision support model for prioritizing projects in a sanitation company (Modelo de apoio à decisão para priorização de projetos em uma empresa de saneamento)	CI-1, CI-2, CI-3
[9]	Portfolio selection of distributed energy generation projects considering uncertainty and project interaction under different enterprise strategic scenarios	CI-1, CI-2, CI-3
[2]	Strategic alignment and value maximization for IT project portfolios	CI-1, CI-2, CI-3
[10]	Managing uncertainty to improve decision-making in NPD portfolio management with a fuzzy expert system	CI-1, CI-2, CI-3
[3]	Project portfolio adjustment and balance: A case study in the chemical sector	CI-1, CI-2, CI-3
[11]	Software project portfolio optimization with advanced multiobjective evolutionary algorithms	CI-1, CI-2, CI-3
[12]	IT project selection model using real option optimization with fuzzy set approach	CI-1, CI-2, CI-3
[5]	Proposal of a prioritizing method for it infrastructure projects	CI-1, CI-2, CI-3
[13]	Multi-criteria and model-based analysis for project selection: An integration of capability-based planning, project portfolio management and enterprise architecture	CI-1, CI-2, CI-3
[14]	Proposal and Solution of a Mixed-Integer Nonlinear Optimization Model That Incorporates Future Preparedness for Project Portfolio Selection	CI-1, CI-2, CI-3

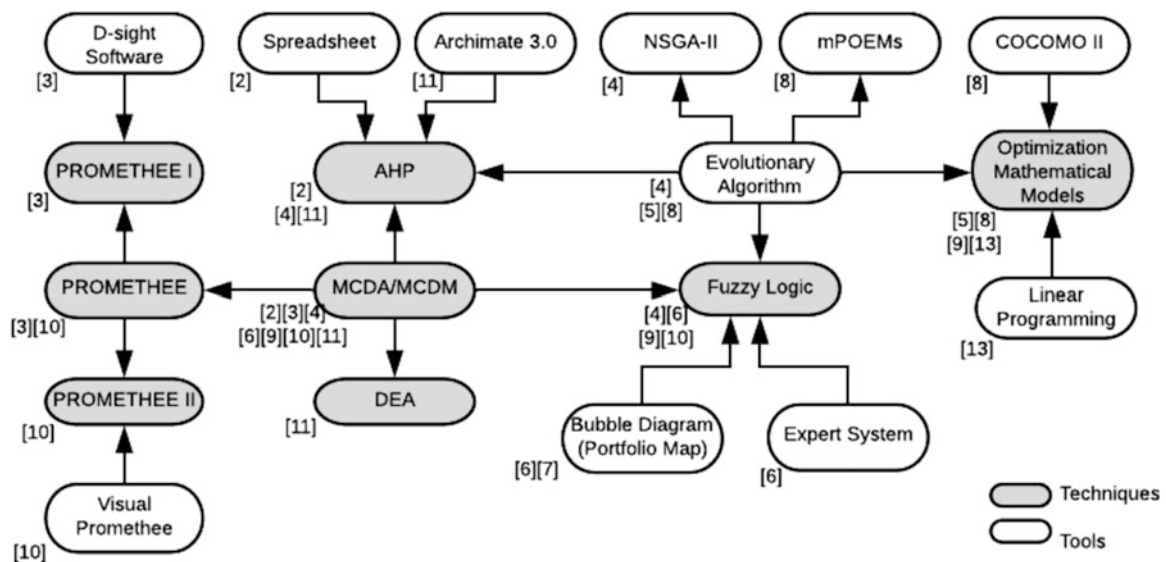


Fig. 25.1 Techniques and tools used in selection and strategic alignment of projects and project portfolio balancing

In a new step, the IEEE search string was changed to apply only to abstracts, so the search returned 20 results. As there was a possibility of false positives, a screening process was necessary to obtain only the articles relevant to the research. The screening process started by reading the abstracts and subsequently applying the exclusion criteria. Based on the reading of the abstracts, 9 articles from IEEE and 2 articles from Scopus were rejected. In some cases, it was not possible for the abstract to identify whether the article was in fact not a false positive and, then, the article was selected to be part of the last stage. In the last step, the articles from the two databases were joined and duplication removed (2 duplications found), resulting in 22 articles. Afterward, the articles were read in full and the inclusion and exclusion criteria were applied. The articles that met the inclusion criteria are shown

in Table 25.1. It is worth mentioning that during the reading of [7], the article [8] was included by snowball, an admissible search technique in systematic mappings. The techniques and tools identified in these articles are shown in Fig. 25.1.

25.2.1.6 Results

In studies that met the inclusion criteria, the use of the Multi-Criteria Decision Analysis (MCDA), also known as Multi-Criteria Decision-Making (MCDM), is notorious. This method presents itself as a viable alternative to assist in the process of prioritizing projects, as it helps in defining the priority of projects that simultaneously deal with conflicting criteria [8].

MCDA can be classified into Multi-Attribute Decision Making (MADM) and Multi-Objective Decision Making

(MODM). In project portfolio selection, MADM can be used to evaluate individual projects, while MODM can be used to optimize the project portfolio [9].

Among the MODM techniques, the use of fuzzy logic, internal type-2 fuzzy numbers (IT2FNs) was observed, which compared to other fuzzy numbers, such as triangular fuzzy numbers and trapezoidal fuzzy numbers, has a higher degree of fuzzification and, therefore, is more suitable for making complex decisions. Weights of the criteria determined by interval type-2 fuzzy analytic hierarchy process (IT2FAHP) technique were integrated into the interval type-2 fuzzy weighted averaging (IT2FWA) operator to obtain the strategic alignment index for projects in [9].

The Analytic Hierarchy Process (AHP) and the Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE) and their variations were also highlighted.

The AHP method was created in the 1970s by Thomas L. Saaty. AHP allows to measure the impact of the different criteria considered in the decision problem in relation to the general objective. This method consists of creating a model that reflects the functioning of the human mind in evaluating alternatives in the face of a complex decision problem, which allows dealing with problems that involve both tangible and intangible values, as it has the ability to work with qualitative variables with based on subjective judgments issued by decision makers [7].

PROMETHEE is considered quite simple, both in design and application, compared to other MCDA methods. The methods of the PROMETHEE family belong to the group of criteria of the overclassing approach. PROMETHEE I provides partial pre-orders and PROMETHEE II provides complete pre-orders [8].

Figures 25.2 and 25.3 graphically translate the identified techniques and the MCDA / MCDM techniques most used in the articles selected in the systematic mapping.

In relation to the tools, the use of evolutionary algorithms is highlighted, as shown in Fig. 25.4. In [9], the non-dominated sorting genetic algorithm-II (NSGA-II) is used to obtain an optimal-Pareto in different scenarios strategic. This algorithm was used due to its lower complexity, fast

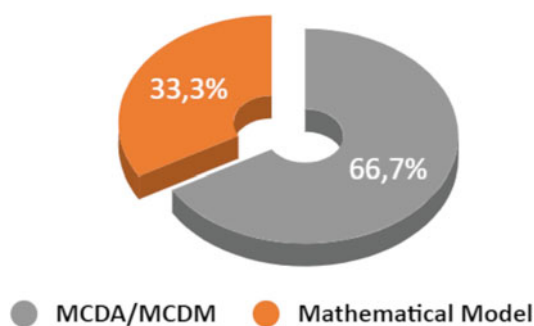


Fig. 25.2 Techniques identified in the selected articles

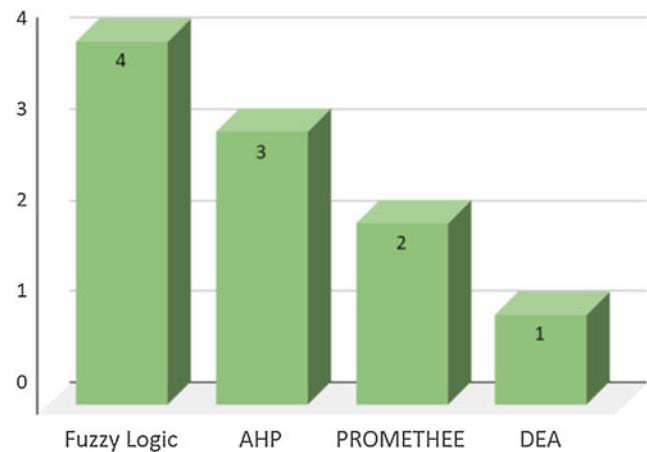


Fig. 25.3 MCDA/MCDM techniques identified and number of articles that used them

execution speed and good convergence of the solution set. In [11], the algorithm mPOEMS is used, which is an iterative optimization algorithm that seeks in each iteration to modify the current solution, called a prototype, which improves its quality. The modifications are represented as sequences of elementary actions (simple mutations in the standard evolutionary algorithms (EAs)), defined specifically for the problem in question. MPOEMS uses an evolutionary algorithm to look for the best action sequence that can be considered an evolved hypermutation. According to [11] mPOEMS produces better or at least competitive results compared to NSGA-II.

25.3 Discussion

The question to be answered with systematic mapping is: What tools and techniques are used (or are they most suitable) for selection and strategic alignment of projects and project portfolio balancing? The articles selected by the inclusion criteria provided us with information that makes it possible to answer this question.

In [7] a model is presented to support the decision-making process in the selection of proposals for scientific research projects in a federal educational institution in a scenario of scarce resources to meet the growing demand for research projects coming from the proponents. In the construction of the model was used the Analytic Hierarchy Process (AHP) method. According to [7], the research results contribute to stimulating the practice of using multi-criteria decision support methods by managers of public and private companies to deal with complex decision problems.

The objective in [8] is to improve project prioritization techniques for managing multiple projects. The study proposes a model that uses the multi-criteria decision support

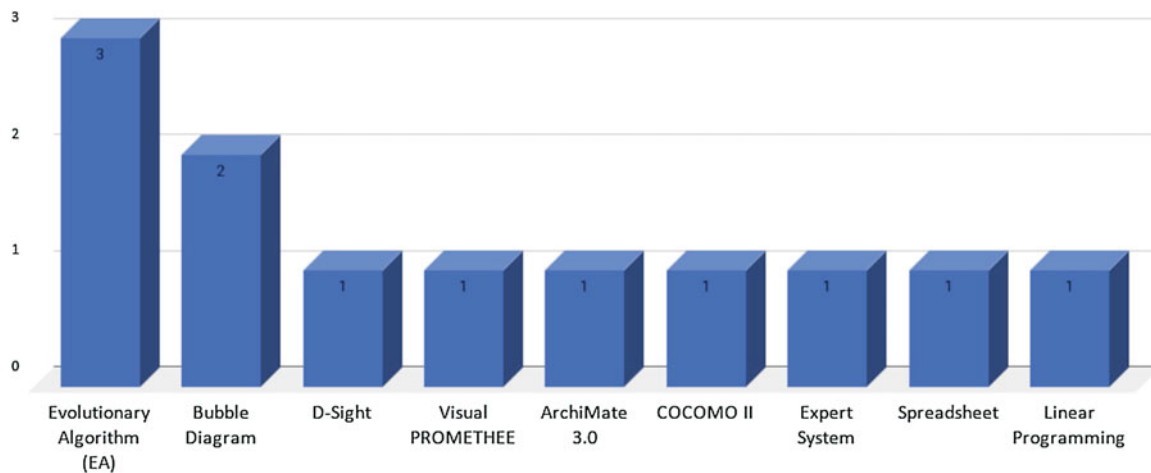


Fig. 25.4 Tools identified in the selected articles and number of articles that used them

method PROMETHEE I to prioritize projects in a Brazilian sanitation and water treatment company, strongly characterized by a hierarchy of low flexibility and which does not yet have the necessary maturity to conduct with your projects. The model used was considered satisfactory, meeting the needs of the company and the manager of projects.

The objective of [9] is to select the ideal portfolio of Distributed Power Generation (DEG) projects in different strategic scenarios, in which the project's uncertainty and interactions are considered. Multi-criteria decision support methods are used. Weights of the criteria are determined using by interval type-2 fuzzy analytic hierarchy process (IT2FAHP = AHP + type 2 interval Fuzzy). Based on strategic interactions, a nonlinear 0–1 programming is formulated, satisfying budgetary constraints and the non-dominated sorting genetic algorithm-II (NSGA-II) is used to obtain the optimal project portfolio. The proposed structure was applied in a case study in Zhejiang province, China. The achievement of strategic objectives was measured by project. According to [9], the results show that the selected portfolios vary according to the strategic objectives of the companies and the study has a practical applied value for project managers in project portfolio management.

In [2] an optimization model is proposed to optimize decision processes for IT portfolios and programs. When formulating the mathematical model, the selection of proposals that are in the portfolio is considered first, then a set of proposal indexes for the projects under review is defined. The study also describes the use of an evolutionary algorithm (EA) to find approximate solutions for the optimization model. According to [2], the EA procedure identifies candidate solutions that exhibit Pareto optimality and also suggests how projects should be assigned to development programs so that the total development time can be reduced [2] was one of the few studies that did not address multi-criteria decision support methods.

[10] proposes a decision-making framework that uses a fuzzy expert system in portfolio management to deal with uncertainty in the development of new products (NPD). The article establishes portfolio assessment models based on fuzzy inference for ambiguous items that are difficult to evaluate numerically. To this end, a specialist system has been developed that facilitates the selection of correct projects to develop balanced investment Research and Development (R&D) programs. The suggested structure was applied to portfolio analysis in an electronics company. The work points out some limitations of the system: (i) the suggested approach focuses more on the evaluation process than on the evaluation criteria and decision-making rules; (ii) the same rule is used regardless of the types of project, however, due to the variety of types of project in a strategic bucket, it is necessary to apply different selection criteria for each type of project; (iii) development focused on two common types of portfolios, so it is necessary to design a portfolio that reflects more versatile standards to balance all aspects of a project.

The objective of [3] is to understand the adjustment step in the context of project portfolio management, highlighting its relationship with the processes of categorization and balancing. The research was developed in a company in the Brazilian chemical sector with data from a thousand projects carried out between 2001 and 2005. Although the company does not use portfolio management methodologies to guarantee the alignment between the business strategy and its project portfolio, the study presents how the adoption of balancing tools allowed to evidence gaps and sources of imbalance in the project portfolio. To analyze the balance of the portfolio, three types of bubble diagrams were elaborated that allowed to show the dynamics of the company's project portfolio. People involved in the case study pointed out as positive points the ease of understanding the diagrams, evidencing the imbalance graphically and pointing out the strengths and weaknesses of the projects in various dimensions.

[11] presents the first phase of implementation a decision support framework focused on the selection of software projects, using an optimization evolutionary algorithm (EA), the mPOEMS. A set of 50 projects that follow the criteria of the COCOMO II model are worked on and the performance of the mPOEMS approach is compared with other multiobjective optimization approaches. The mPOEMS performed significantly better than NSGA-II and SPEA2, state-of-the-art multiobjective optimization evolutionary approaches. According to [11], this optimization framework is able to find efficient portfolios on or close to the Pareto-optimal front.

[12] presents a mathematical fuzzy model to optimize the option value for a multistage portfolio of IT projects. The goal is to develop a fuzzy portfolio selection model to optimize the IT portfolio for risk-averse decision makers in an uncertain environment. The IT portfolio selection problem is formulated as a fuzzy programming model with a variation between 0 and 1, which can deal with both uncertain and flexible parameters, to determine the optimal (or ideal) project portfolio. The idea of fuzzy optimization of the option value of the portfolio is to maximize the overall value and minimize the risk of falling of the portfolio selected for financing. A transformation method based on the theory of qualitative possibility is developed to convert the fuzzy model of portfolio selection into a clear mathematical model from a risk-averse perspective. The transformed model can be solved by an optimization technique. According to [12], the optimization model and the solution approach can help IT managers in making optimal financing decisions for project prioritization.

[5] presents a project prioritization method, which contemplates the application of a quality matrix adapted from the Fuzzy QFD technique and the MCDA PROMETHEE II method for the selection and prioritization of portfolio projects, in a scenario of uncertainty and resource limitation. The method was applied in the portfolio management process of Information Technology infrastructure projects at the Army Integrated Telematics Center (CITEx). According to [5] the proposed method resulted in an ordering consistent with the current investment scenario of the organization and with the empirical analysis of the specialists. Despite the prioritization process being based on subjective evaluations by specialists, the analysis of projects according to previously defined criteria and benefits and the application of quality tools and multicriteria analysis, addressed the uncertainties and minimized the risks inherent to the decision-making process.

[13] proposes a method that aims to assist in making organizational decisions in relation to investments based on capacities and multiple selection criteria. The method combines the techniques Analytical Hierarchy Process (AHP) and Data Envelopment Analysis (DEA), and several analyzes

based on the Enterprise Architecture (EA) Model, structured in an eight-step iterative process. The case study was carried out based on a case described in the literature, a major European Energy Supplier (called in the study by EES). In addition, the method was evaluated through a panel with 5 experts. The results suggest that the method has the potential to be used in practice, provided sufficient guidance is provided [13] reports some limitations, including: (i) the method includes only the AHP and DEA techniques, while there are several other techniques suitable for project selection; (ii) more comprehensive guidelines must be provided for the use of the techniques included in the method (for example, analysis of benefits and impacts); (iii) the need for further validation and evaluation studies on the method to improve the generalization of results is recognized.

[14] presents a mathematical optimization model, implemented by a linear programming algorithm, for portfolio selection that considers four main performance measures for project management: (i) maximizing value, (ii) strategic alignment, (iii) balance and (iv) future preparation. The model was tested with real data from a start-up that operates in the green mobility sector in Canada and with a Brazilian company that develops educational software. The model works with several parameters, among them, total financial resources, maximum and minimum number of projects in the portfolio and number of organizational objectives. The objective of the proposed mathematical function is to maximize the revenue generated by the portfolio, given by the difference between the total expected production of the portfolio and the total input invested in the portfolio. According to [14], the results obtained with the model are consistent with the results obtained in practice by companies without using it.

As can be seen, the Multicriteria Decision Support Methods (MCDA/MCMD) are presented in the literature as the most used technique to assist managers in the selection of projects from a portfolio, in more complex scenarios. This is due to the fact that these methods consider more than one aspect of the problem and, therefore, make it possible to evaluate actions according to a set of criteria.

The studies also presented proposals for mathematical models and expert systems that used fuzzy logic, which is also an MCDA/MCDM technique. Fuzzy logic is a formal mathematical theory for the representation of uncertainties and is based on the fact that the existing sets in the real world have no precise limits. The theory of fuzzy sets has emerged as an effective technique for describing uncertainties and has been widely used in the selection of project portfolios [9].

To work with the models and techniques, the following tools were used: (i) electronic data spreadsheet, (ii) D-sight software, (iii) Visual PROMETHEE software, (iv) Archi-Mate 3.0 software, (v) algorithm of linear programming, (vi)

evolutionary algorithms (EA), such as mPOEMs and the non-dominated genetic classification algorithm II (NSGAI), (vii) bubble diagrams, (viii) a specialist system and, furthermore, (ix) the COCOMO II model.

The variation in techniques and tools is explained by the fact that there are variations in the types of projects. Each project can be better managed by a specific technique depending on the context [15] states that “the use of certain methods and criteria leads to specific performance results, revealing that the portfolio’s performance depends on the methods and criteria used for project selection”. Still in [15], “empirical evidence of the link between methods and results is rare and more research is needed to fully understand the relationships and links between portfolio management methods and performance indicators”.

In the studies of this systematic mapping, performance indicators were not presented to portray the results of before and after the application of the techniques and tools adopted for the selection and strategic alignment of projects and balancing the project portfolio. The studies also do not present comparative data on the application of different techniques to the same set of projects. It was therefore not possible to ascertain among the techniques and tools used which are the most appropriate.

25.4 Conclusion

This work described a systematic mapping carried out with the purpose of identifying which tools and techniques are used (or are more suitable) for strategic project selection and alignment and project portfolio balancing.

The research was conducted from the digital libraries Scopus and IEEE and allowed to verify that the vast majority of the techniques used are Multicriteria Decision Support Methods (MCDA/MCDM) and mathematical models that make use of fuzzy logic, in addition to evolutionary algorithms as main tool.

Several tools were used to work with the models and techniques. The variation in techniques and tools is explained by the fact that different techniques lead to different results, which implies that the performance of the portfolio depends on the methods and criteria used in the selection of projects.

In the studies of this mapping, performance indicators were not presented to portray the results of before and after the application of the techniques and tools used. The studies also do not present comparative data on the application of different techniques to the same set of projects. It was therefore not possible to ascertain among the techniques and tools used which are the most appropriate. However, the systematic mapping of the literature made it possible to identify which are the most used.

References

1. A.D. Souza, A.R.C. Rocha, D.C.S. Santos, A proposal for the improvement of project’s cost predictability using earned value management and historical data of cost – an empirical study. *Int. J. Software Eng. Knowl. Eng.* **25**(01), 27–50 (2015). <https://doi.org/10.1142/s0218194015400021>
2. I. Robert Chiang, M.A. Nunez, Strategic alignment and value maximization for IT project portfolios. *Inf. Technol. Manag.* **14**(2), 143–157 (2013). <https://doi.org/10.1007/s10799-012-0126-9>. ISSN: 1385951X
3. M. Padovani, M.M. Carvalho, A.R. Muscat, Project portfolio adjustment and balance: a case study in the chemical sector. *Produção* **22**(4), 674–695 (2012). <https://doi.org/10.1590/S0103-65132012005000064>. ISSN: 19805411.
4. Project Management Institute (PMI), *A guide to the project management body of knowledge (Guia PMBOK®)*, 5th edn. (Project Management Institute, Inc., Pennsylvania, 2013)
5. P. de Araújo Farias, C.M. de Lima, S.B.S. Monteiro, A.C.B. Reis, Proposal of a prioritizing method for it infrastructure projects. *RISTI* **2020**(E27), 763–776 (2020). ISSN: 16469895.
6. C. Wohlin, P. Runeson, M. Host, M.C. Ohlsson, B. Regnell, A. Wesslen, *Experimentation in software engineering. Capitulo 4: Systematic Literature Review*. ISBN 978-3-642-29043-5 ISBN 978-3-642-29044-2 (eBook). DOI <https://doi.org/10.1007/978-3-642-29044-2>. Springer Heidelberg New York Dordrecht London. Library of Congress Control Number: 2012940660
7. M.C. Ribeiro, A.D. Alves, The problem of research project portfolio selection in educational organizations: a case study. *Gest. Prod., São Carlos* **24**(1), 25–39 (2017). <https://doi.org/10.1590/0104-530X2089-16>. ISSN: 0104530X
8. M.T. Lima, E.C. Oliveira, L.H. Alencar, Modelo de apoio à decisão para priorização de projetos em uma empresa de saneamento. *Production* **24**(2), 351–363 (2014). <https://doi.org/10.1590/S0103-65132013005000072>
9. Y. Wu, C. Xu, Y. Ke, X. Li, L. Li, Portfolio selection of distributed energy generation projects considering uncertainty and project interaction under different enterprise strategic scenarios. *Appl. Energy* **236**, 444–464 (2019). <https://doi.org/10.1016/j.apenergy.2018.12.009>. ISSN: 03062619
10. J. Oh, J. Yang, S. Lee, Managing uncertainty to improve decision-making in NPD portfolio management with a fuzzy expert system. *Expert Syst. Appl.* **39**(10), 9868–9885 (2012). <https://doi.org/10.1016/j.eswa.2012.02.164>. ISSN: 09574174.
11. T. Kremmel, J. Kubalík, S. Biffl, Software project portfolio optimization with advanced multiobjective evolutionary algorithms. *Appl. Soft Comput. J.* **11**(1), 1416–1426 (2011). <https://doi.org/10.1016/j.asoc.2010.04.013>. ISSN: 15684946.
12. S. Pushkar, A. Mishra, IT project selection model using real option optimization with fuzzy set approach. *Commun. Comput. Inf. Sci.* **194**, 116–128 (2011). https://doi.org/10.1007/978-3-642-22603-8_12. ISSN: 18650929.
13. A. Aldea, M.E. Iacob, M. Daneva, L.H. Masyhur, Multi-criteria and model-based analysis for project selection: An integration of capability-based planning, project portfolio management and enterprise architecture, in *2019 IEEE 23rd International Enterprise Distributed Object Computing Workshop (EDOCW)*, (2019). <https://doi.org/10.1109/EDOCW.2019.00032>. ISSN: 15417719.
14. T.C.L. Albano, E.C. Baptista, F. Armellini, D. Jugend, E.M. Soler, Proposal and solution of a mixed-integer nonlinear optimization model that incorporates future preparedness for project portfolio selection, in *2019. IEEE Trans. Eng. Manage.*, (2019). <https://doi.org/10.1109/TEM.2019.2920331>. ISSN: 00189391.
15. M. Lerch, P. Spieth, Innovation project portfolio management: a qualitative analysis. *IEEE Trans. Eng. Manage.* **60**(1), 18–29 (2013). <https://doi.org/10.1109/TEM.2012.2201723>. ISSN: 0018-9391.

Methods for Detecting Fraud in Civil and Military Service Examinations: A Systematic Mapping

Roberto Paulo Moreira Nunes, Rodrigo Bonacin, and Ferrucio de Franco Rosa

Abstract

Civil and military service examinations are carried out in several countries for the recruitment and admission of public servants in various spheres/levels of government. This is considered an effective and rational method for selection based on merit. Due to the constant economic variations and the stability provided by public offices, the interest in some offered positions can be huge. Criminals specialized in defrauding public examinations offer candidates the possibility of facilitated and illegal admission. Various types of information could be submitted to methods and techniques (e.g., application data, test performance, geodata, etc.) to detect fraud. We present a systematic mapping of the literature on fraud detection methods in several domains, which can be adapted and improved to detect fraud in public examinations. 31 articles were identified, and after analysis, 19 selected works were analyzed and classified. The usages of machine learning and data mining techniques were uppermost methods adopted in the analyzed papers. This work is aimed at researchers who seek to develop fraud detection techniques in admission exams.

Keywords

Detection · Fraud · Civil service examination · Machine learning · Data mining

R. P. M. Nunes (✉)
University of Campo Limpo Paulista (UNIFACCAMP), Campo Limpo Paulista, SP, Brazil

R. Bonacin · F. de Franco Rosa
University of Campo Limpo Paulista (UNIFACCAMP), Campo Limpo Paulista, SP, Brazil

Renato Archer Information Technology Center (CTI), Campinas, SP, Brazil

26.1 Introduction

Civil Service is a term that represents the government, composed mainly of career employees hired by professional merit criteria (e.g., curriculum and tests). Civil (or Public) servant is the professional who works in a government department or agency. Hiring methods, as well as details about how these methods are used, vary according to the country. Civil (and Military) service examinations are competitions (or tests) carried out in several countries for the recruitment and admission of public servants in various spheres/levels of government (e.g., federal, state, and municipal). Examination based methods are considered effective and rational alternatives for recruiting civil servants using merit based criteria. Some positions attract huge number of candidates due to stability and provided benefits. Worldwide, important vacancies are contested by highly qualified professionals every year, who are subjected to complex tests. Indeed, frequently there is an ecosystem around civil service examinations involving various stakeholders, such as: organizers, candidates, inspectors, schools and preparatory courses, websites and publishers, among others.

Criminals specialized in defrauding public examinations offer to candidates the possibility of facilitated and illegal admission. Frauds can be carried out in various ways, such as in the manipulation of results, in the previous disclosure of templates, in direction to carry out tests, or even in the use of electronic points to transmit responses to candidates [1]. The amount charged by gangs is directly proportional to the importance of the desired position and to the employed technique.

Several institutions that organize the competitions have datasets, which could be used to inform methods and techniques for detecting fraud. These datasets include various types of information, such as biographical and geographical

information about each candidate, performance in the intellectual test and other selection phases, as well as information about previous participations.

We present a systematic mapping of the literature with the following objectives: (i) to survey methods and techniques that represent the state-of-the-art in fraud detection; and (ii) to point out gaps in the literature and to discuss research challenges.

The remainder of this article is organized as follows: Section 2 presents a conceptual background from a literature review with the same purpose as our review (i.e., related work); Sect. 3 describes the methodology used to carry out the review and analysis of the articles; Sect. 4 details the results categorized according to the techniques adopted in each work; and Sect. 5 presents a discussion on the results and the conclusions.

26.2 Background and Related Work

Reference [2] was considered a related work because it presents objectives close to ours. This work focuses on surveying works and techniques for preventing or detecting fraud in credit card transactions, both in online and in-person purchases. This reference also provides background on some adopted domain concepts in our work.

Prevention is pointed out as the most important objective, since it occurs before the transaction takes place, unlike detection, which works at a later stage. Historical data and user behavior patterns are used to check and verify whether a transaction is fraudulent or not. Detection response time is an important factor, especially in online transactions. In the survey, the types of credit card fraud were highlighted: (i) card not present in Internet purchases; (ii) skimming (physical device camouflaged in card readers); (iii) phishing (establishing a trust relationship to obtain confidential data through fake websites, for example); (iv) lost (or stolen) card used before blocking by the user.

Approaches to data mining have been identified in [2], with the following objectives: (i) *Classification*: differentiate several categories of objects by using a model. The classification provides the object labels; the labels are predefined and distinct. (ii) *Grouping*: objects are divided into conceptually significant groups called clusters. In the same cluster, the objects are very similar in terms of resource. (iii) *Prevision*: used to foresee the continuous value. Based on historical data, patterns are created and numerical values estimated. (iv) *Outlier detection*: data objects that are completely different from the entire remaining dataset are identified. Outlier detection means measuring the “distance” between the data points and the outlier object. Detecting extreme values is a crucial issue in the data mining field of research. (v) *Regression*: shows the relationship between more than one dependent and

independent variable. Regression is one of the best statistical methods, representing a benchmarking.

Techniques based on statistics and computing are also identified in [2], namely: (i) Artificial immune system; (ii) Bayesian Network; (iii) Logistic Regression; (iv) Neural Network; (v) Support Vectors Machine; (vi) Genetic Algorithm; (vii) Decision Tree; (viii) Self-organized map; and (ix) Hybrid methods. The authors point out that Machine Learning (ML) techniques are the most used in fraud detection, which achieve a good accuracy and detection rate. Our systematic mapping differs from the described review in that: (1) it does not focus exclusively on credit card fraud; (2) does not exclusively address techniques based on ML (e.g., data mining, statistics, processes).

26.3 Review Methodology

Our systematic mapping of the literature is based, with adaptations, on the Kitchenham method [3]. The review process, including the activities and the produced documents, is briefly presented in Fig. 26.1. The proposed process consists of 3 phases, as follows:

Planning was carried out based on the review protocol, containing the theme and objectives of the review. The following scientific databases were selected, once they cover relevant research on the topic: ACM Digital Library;¹ IEEE Xplore;² Google Scholar³ e Springer Link.⁴ Inclusion criteria were defined: (1) Papers from journals or proceedings of scientific events, which the full text is available in the scientific databases; (2) Published after 2010; (3) Works that present methods and techniques for detecting fraud; (4) Works published in the English language. Exclusion criteria were defined: (1) Research area other than Computer Science; (2) Works that not focus on topics related to this review; (3) Short papers and Abstracts. Keywords and the search string were defined, namely: “(method OR procedure OR architecture) AND (detection OR evaluation OR assessment) AND (contest OR competition OR credit OR card) AND Fraud”. During **Conduction**, the search string was adapted, according to the particularities of the search engines (Table 26.1); specifically, the form of insertion of keywords, considered fields and refinement in searches. In case of a null return of articles, the keywords were adjusted to a lesser restriction. Date constraints: Articles published between 2010 and 2020; and most recent literature reviews, published between 2018 and 2020.

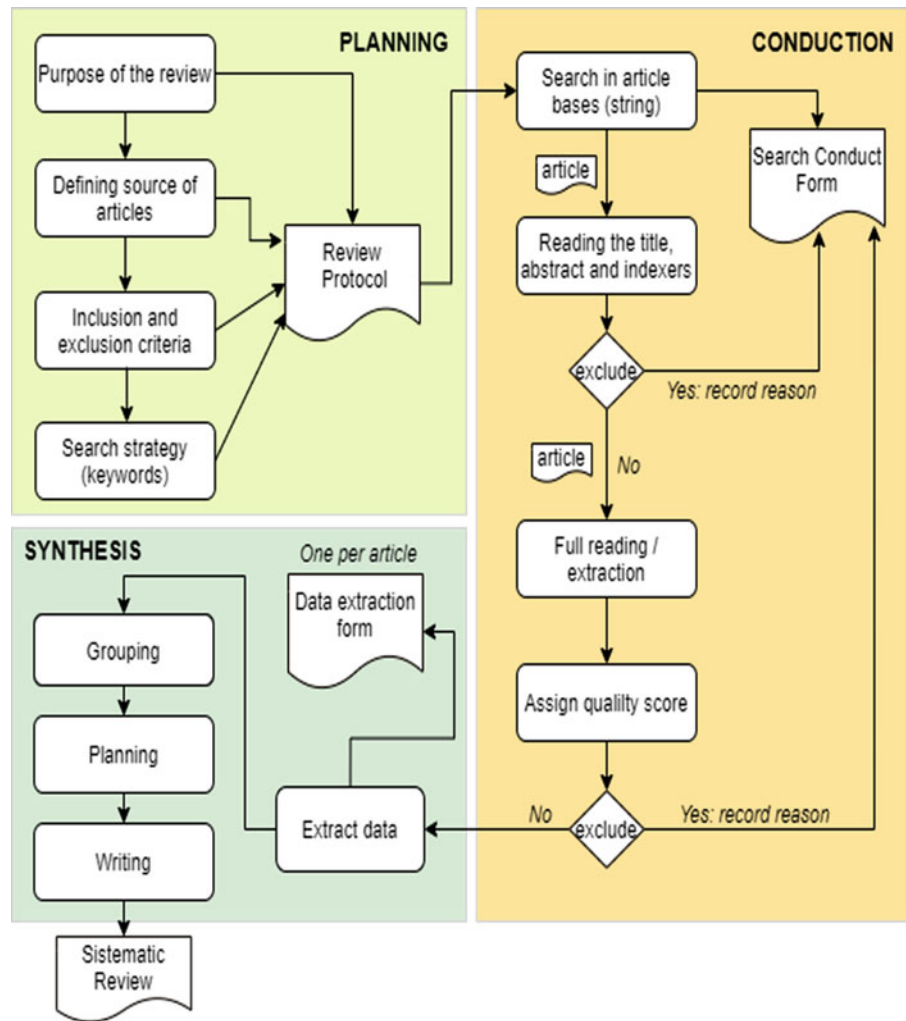
¹<http://portal.acm.org/>

²<http://ieeexplore.ieee.org>

³<http://scholar.google.com>

⁴<http://links.springer.com>

Fig. 26.1 Process of systematic review



5 studies cited by the analyzed studies (Snowballing⁵), in such case, even when outside the periods, they were included in the review. An auxiliary spreadsheet and Mendeley⁶ were used in the conduction and synthesis phases. We performed a complete reading of the selected works. No article was added or deleted based on research quality criteria. The *Synthesis* phase was performed based on the review protocol [3].

26.4 Review Results

We present an analysis of the 19 selected studies according to five categories as follows: Subject. 4.1 presents solutions based on ML; Subject. 4.2 presents solutions based on data mining; and Subject. 4.3 presents solutions grounded on statistical techniques; Subject. 4.4 presents research concerning

detection of fraud in multiple choice exams; and Subject. 4.5 presents research regarding fraud in public tenders.

26.4.1 Solutions Based on Machine Learning

An approach for detecting fraud in online auctions is proposed in [4]. The objective is to reduce the quantity of the attributes evaluated to describe the characteristics of the participants in this sales modality. The sellers' behavior is evaluated, by considering only the fifth end of the available history, thus reducing the computational efforts. By reducing transaction histories, the total cost of detection can be greatly reduced, while maintaining reasonable detection accuracy. The model built to validate the study considered several data mining techniques, namely: C.45, Cart, Ada Boost, Naive Bayes, NB Trees and Logistic Regression.

From applying the techniques, an accuracy of 91–95% was obtained in the detection of the built late-profiling models. An approach for detecting fraud in credit card purchases is proposed in [5]. Features are extracted, based on the fre-

⁵*Snowballing*. The snowballing technique consists of using certain articles that had been cited by papers that were included in the review.

⁶<http://www.mendeley.com>

Table 26.1 Search strings and collected articles

Scientific Databases and Search Strings	Collected (included)
<i>IEEE Xplore</i> (2010–2020): ((“Document Title”: “method” OR “Document Title”: “procedure” OR “Document Title”: “architecture”) AND (“Document Title”: “detection” OR “Document Title”: “evaluation” OR “Document Title”: “assessment”) AND (“Document Title”: “fraud”)).	16 (9)
<i>IEEE Xplore</i> (2018–2020): (((“Document Title”: detection) AND “Document Title”: fraud) AND “Document Title”: survey).	1 (1)
<i>Springer Link</i> (2010–2020): field “where the title contains “: “detection fraud”.	1 (1)
<i>Google Scholar</i> (2010–2020): allintitle: method OR process “detection fraud”.	6 (2)
<i>ACM Digital Library</i> (2010–2020): [[Publication Title: “method”] OR [Publication Title: “procedure”] OR [Publication Title: “architecture”]] AND [[Publication Title: “detection”] OR [Publication Title: “evaluation”] OR [Publication Title: “assessment”]] AND [Publication Title: “fraud”]	7 (2)

quency of user transactions and rules of individual and group behavior, to classify transactions as legitimate or fraudulent. To validate the proposal, the Random Forest technique was used as a binary classifier. Reference [6] presents a solution to increase precision in detecting fraud of users with low frequency of transactions, i.e., users who do not allow the creation of a more accurate individual profile. A DBSCAN clustering algorithm is used to assign the user to a group of similar users. After that, by means of a Naive Bayes algorithm, it checks if a transaction is fraudulent, considering the private and group environment. Reference [7] presents a method that considers analyzes in subspaces defined by two or three variables recorded in transactions. Of these subspaces, transaction speed and acceleration are estimated as input vectors for a classifying process. The linear and quadratic discriminant analysis and random forest are implemented as classifiers. The classification results obtained for each subspace are then merged to obtain an overall result by using the alpha integration algorithm. Neural networks are used to build financial fraud detection models. However, the accuracy of the models decreases with time when they are deployed in online detection systems. Reference [8] proposes a method of Incremental Virtual Learning (IVL) aiming to continuously update neural networks, maintaining thus the performance of the model when labels of new transactions are not available. Reference [9] proposes a federation-based method for detecting credit card frauds. The decentralized use of ML in the data of each institution of the federation is proposed. The learning model is then shared with the other federated entities, without exposing their transaction data. Reference [10] presents a method based on the label propaga-

tion algorithm to extract resources from the related network. From the related network that contains known fraudulent users and the relationship between them, a custom label propagation algorithm is used to infer the likelihood of fraud by the unknown user. Reference [11] addresses the use of big data in conjunction with ML to identify fraud in companies' balance sheets.

26.4.2 Solutions Based on Data Mining

A method of clustering user environments based on unbalanced data for detecting credit card fraud is proposed in [12]. The authors propose to divide the user environment into several groups of environments by using k-means, remove noise, and rank the sample. Reference [13] presents a method for detecting fraud in the use of credit cards. It uses the k-means cluster and then finds discrepancies in the resulting clusters by using HMM (Hidden Markov Method). The algorithm effectively divides the numbers into clusters and then detects outliers; the credit card number is validated by using the Luhn algorithm.⁷ Reference [14] proposes a framework for fraud detection in credit card transactions. The framework consists of the following components: (i) a rule-based component, which uses a decision tree (supervised); (ii) a component that performs a trend analysis, with calculation of dissimilarities (semi-supervised); and (iii) a scenario-based component, where the extent of similarities in the sequence of transactions with the known fraud scenarios is calculated.

26.4.3 Solutions Based on Statistics

Reference [15] proposes a fraud detection method, which measures a value of fraud risk for a given m-banking transaction. Unlike existing methods, which usually assume that the different risk factors for marginal fraud are independent of each other, the proposed method could capture evasive fraud patterns caused by fraud risk factors that are dependent on or independent of each other. Reference [16] proposes a method of detecting fraud in online transactions, based on individual environments. By considering multiple dimensions of historical transaction records, a user transaction environment is generated. Then, an algorithm is proposed to determine the optimal risk limit for each user. Finally, combining the transaction environment and the ideal risk limit, a benchmark of the user environment is formed, which will be used to build the multidimensional hypersphere model. The transactions are adjusted to this hypersphere, showing whether new transactions are normal or fraudulent. Reference [17]

⁷The Luhn algorithm is a checksum formula that is frequently used to validate credit card numbers [23].

proposes a signature-based method for detecting frauds in electronic commerce. The proposed signature is defined by a set of attributes that receive a set of variables related to a user's behavior in an e-commerce environment. The signature points out behavior deviations in relation to the user's recent activity, allowing us to detect potential fraud situations.

26.4.4 Frauds in Multiple Choice Exams

Three researches focusing on methods for detecting fraud in multiple choice exams was identified. Reference [18] proposes a method for detecting and evaluating cheating in university admission exams by using supervised classification. Reference [19] uses data mining and Bonferroni's correction to detect excessive similarity in multiple-choice responses between pairs of candidates. Reference [20] applied data mining algorithms and statistical tools (hierarchical clustering, dendrogram tree and question difficulty weights) to determine similarities between answers. A visual analysis is also performed, using a heat map, to identify patterns in the exam scores.

26.4.5 Frauds in Public Tenders

In the context of public tenders, the following works were identified. Reference [21] proposes to use a probabilistic ontology, by means of Probabilistic OWL (PR-OWL), to automate fraud detection in public purchases. Reference [22] proposes a model that aims to extract information from big data related to governmental procurement. Aiming to identify cartel formation (fraudulent association between companies) in public procurement, a software tool was created using data mining techniques (grouping and association rules) in conjunction with a multi-agent approach to. Table 26.2 presents a summary of the analyzed works, categorized by techniques and application domains. In general, the prevalent application domains were related to financial transactions, especially those carried out online by using credit cards.

26.5 Discussion E Conclusion

The selection of the best candidates for governmental institutions is essential for the good functioning of their services. Thus, the integrity of the selection process is crucial, seeking to prevent or mitigate the occurrence of frauds.

We have presented a systematic mapping of the literature on fraud detection in several domains. Following our review protocol, studies that propose specific methods for the target domain (selection exams of public servants) were not found

Table 26.2 Synthesis of analyzed works

Refs.	Technique			Application Domain								
	ML	DM	ST	1	2	3	4	5	6	7	8	9
[4]	X							X				
[5]		X								X		
[6]	X				X							
[7]	X									X		
[8]	X								X			
[9]	X									X		
[10]	X					X						
[11]	X			X								
[12]	X									X		
[13]		X								X		
[14]		X			X							
[15]			X				X					
[16]			X						X			
[17]			X		X							
[18]		X									X	
[19]		X									X	
[20]		X	X								X	
[21]			X									X
[22]		X	X									X

Technique: (ML) Machine Learning. (DM) Data Mining. (ST) Statistics. **Application Domain:** (1) Accounting. (2) E-commerce. (3) Financial/Loans. (4) Internet Banking. (5) Online auction. (6) Financial Transactions. (7) Credit Card Transactions. (8) Multiple Choice Exams. (9) Public Tenders

in the scientific bases; during the snowballing process we have found works describing frauds in the e-Gov context (e.g., frauds in university entrance exams, or in public procurement bids). This points out that there are open issues to be addressed, including the proposition of novel methods and techniques for (semi)automatic detection of frauds.

The studies were analyzed and synthesized according to their approaches and techniques. Most of the works present methods for detecting fraud in the financial domain; specifically, frauds related to transactions over the Internet, to e-commerce platforms, or to credit cards. The selected articles mostly use data mining and ML techniques. Techniques identified in this review are promising to support fraud detection methods in civil or military service examinations.

References

1. J.C. Januário, Fraude em concurso público (2009) [Online]. Available: <http://repositorio.ucpparana.edu.br/index.php/direito/article/view/29/30>.
2. R.R. Popat, J. Chaudhary, A survey on credit card fraud detection using machine learning, Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI 2018, 45, 13, 1120–1125, 2018, <https://doi.org/10.1109/ICOEI.2018.8553963>.

3. B. Kitchenham, Procedures for performing systematic literature reviews, *Jt. Tech. Report*, Keele Univ. TR/SE-0401 NICTA TR-0400011T.1, 33, TR/SE-0401, 33, 2004
4. J.S. Chang, W.H. Chang, A cost-effective method for early fraud detection in online auctions, in *International Conference on ICT and Knowledge Engineering*, (2012), pp. 182–188. <https://doi.org/10.1109/ICTKE.2012.6408551>
5. Y. Xie, G. Liu, R. Cao, Z. Li, C. Yan, C. Jiang, A feature extraction method for credit card fraud detection, in *Proceedings – 2019 2nd International Conference on Intelligent Autonomous Systems, ICoIAS 2019*, (2019), pp. 70–75. <https://doi.org/10.1109/ICoIAS.2019.00019>.
6. Z. Zhang, L. Chen, Q. Liu, P. Wang, A fraud detection method for low-frequency transaction. *IEEE Access* **8**, 25210–25220 (2020). <https://doi.org/10.1109/ACCESS.2020.2970614>.
7. A. Salazar, G. Safont, L. Vergara, A new method for fraud detection in credit cards based on transaction dynamics in subspaces, in *Proceedings – 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, (2019), pp. 722–725. <https://doi.org/10.1109/CSCI49370.2019.00137>.
8. T. Ma et al., An unsupervised incremental virtual learning method for financial fraud detection, in *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, vol. 2019, (2019). <https://doi.org/10.1109/AICCSA47632.2019.9035259>
9. W. Yang, Y. Zhang, K. Ye, L. Li, FFD: a federated learning based method for credit card fraud detection, in *Big Data – BigData 2019. Lecture Notes in Computer Science*, ed. by K. Chen, S. Seshadri, L. J. Zhang, vol. 11514, (2019), pp. 18–32
10. P. Zhao, X. Fu, W. Wu, D. Li, J. Li, Network-based feature extraction method for fraud detection via label propagation, in *2019 IEEE International Conference on Big Data and Smart Computing, Big-Comp 2019 – Proceedings*, vol. 1, (2019). <https://doi.org/10.1109/BIGCOMP.2019.8679414>
11. Y.J. Chen, C.H. Wu, On big data-based fraud detection method for financial statements of business groups, in *Proceedings – 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*, (2017), pp. 986–987. <https://doi.org/10.1109/IIAI-AAI.2017.13>.
12. Q. Li, Y. Xie, A behavior-cluster based imbalanced classification method for credit card fraud detection. *ACM Int. Conf. Proc. Ser.*, 134–139 (2019). <https://doi.org/10.1145/3352411.3352433>
13. P. Bhati, M. Sharma, Credit card number fraud detection using K-means with hidden markov method. *SSRG2*(3), 104–108 (2015)
14. A. Eshghi, M. Kargari, Introducing a method for combining supervised and semi-supervised methods in fraud detection, in *Proceedings of 2019 15th Iran International Industrial Engineering Conference, IIIEC 2019*, (2019), pp. 23–30. <https://doi.org/10.1109/IIIIEC.2019.8720642>.
15. A.A.I. Alnajem, N. Zhang, A copula-based fraud detection (CFD) method for detecting evasive fraud patterns in a corporate mobile banking context, in *2013 International Conference on IT Convergence and Security, ICITCS 2013*, (2013), pp. 0–3. <https://doi.org/10.1109/ICITCS.2013.6717772>.
16. L. Chen, Z. Zhang, Q. Liu, L. Yang, Y. Meng, P. Wang, A method for online transaction fraud detection based on individual behavior, in *ACM Int. Conf. Proceeding Ser.*, (2019). <https://doi.org/10.1145/3321408.3326647>.
17. O. Belo, G. Mota, J. Fernandes, A signature based method for fraud detection on e-commerce scenarios, in *Analysis of Large and Complex Data*, (Springer, Cham, 2016), pp. 497–506
18. E.R. Cavalcanti, C.E. Pires, E.P. Cavalcanti, V.F. Pires, Detection and evaluation of cheating on college exams using supervised classification. *Inf. Educ.* **11**(2), 169–190 (2012). <https://doi.org/10.15388/infedu.2012.09>
19. G.O. Wesolowsky, Detecting excessive similarity in answers on multiple choice exams. *J. Appl. Stat.* **27**(7), 909–921 (2000). <https://doi.org/10.1080/02664760050120588>.
20. M. Chen, Detect multiple choice exam cheating pattern by applying multivariate statistics. *Proc. Int. Conf. Ind. Eng. Oper. Manag.* **2017**, 173–181 (2017)
21. R.N. Carvalho, K.B. Laskey, P.C.G. Costa, M. Ladeira, L.L. Santos, S. Matsumoto, Probabilistic ontology and knowledge fusion for procurement fraud detection in Brazil. *CEUR Workshop Proc.* **527**, 3–14 (2009). https://doi.org/10.1007/978-3-642-35975-0_2
22. S.S.C.V. Ralha, C. Ghedini, A multi-agent data mining system for cartel detection in Brazilian government procurement. *Expert Syst. Appl.* **39**(14), 11642–11656 (2012). <https://doi.org/10.1016/j.eswa.2012.04.037>.
23. J. Chambers, T. Robison, D. Dorsner, S. Manickam, D. Konisky, Luhn validation and data security across multiple active domains, U.S. Patent n. 8,812,844, 2014.

Rafael Leite, Adler Diniz, and Melise De Paula

Abstract

Although the term Smart City has a wide range of definitions, it is a consensus in the bibliography that public participation is essential. However, cities are not always successful in ensuring citizen involvement, which can compromise governance efficiency. This work presents a Literature Systematic Mapping that aims to understand what are the aspects that influence citizens' interest to engage in public participation policies, how these policies are being implemented, and how citizens' engagement affects the governance of a smart city. To accomplish these goals, 33 academic papers were selected through a rigorous search protocol.

Keywords

Smart cities · Public participation · Systematic literature mapping · Citizens engagement · ICTs

be heard, but, the effort to include the population in this type of governance can be considered only symbolic since decision making is still controlled by government officials and Citizen Power, where the population has increasing degrees of decision-making clout.

In past decades, with the popularization of the internet and the emergence of Information and Communication Technologies (ICTs), new inclusive ways to guarantee people's participation in political decisions have emerged and progressively more locations are making efforts to increase citizens' involvement in political decisions. The term "Smart City" appeared without a clear definition, but usually refers to cities with high use of technological resources to promote development based on the interaction between residents and public administration. One of the most accepted descriptions of the term, characterizes a "Smart City" as a city well performing in a forward-looking way in six attributes: Economy, People, Governance, Mobility, Environment, and Living [2].

This work explores the relevant points about public participation in smart city projects with a Systematic Literature Mapping.

27.1 Introduction

Citizen involvement in public administration is desirable in democracies and is not a recent issue. Still in the 1960s [1], analyzed the participatory interactions between government and population, identifying three different levels of participation: Nonparticipation, where all decisions are made by the powerholders; Tokenism, where the population can hear and

R. Leite (✉)

Universidade Federal de Itajubá, Itajubá, Brazil
e-mail: rafaelnleite@unifei.edu.br

A. Diniz · M. De Paula

Institute of Systems Engineering and Information Technology,
Universidade Federal de Itajubá, Itajubá, Brazil
e-mail: adlerdiniz@unifei.edu.br; melise@unifei.edu.br

27.2 Methodology

A Literature Systematic Mapping was conducted aiming to better understand the role of citizens in smart city projects. To reduce biases and establish technical and replicable criteria, it was necessary to adopt a sequence of strict and pre-established procedures to be followed. This plan of action was used to conduct research for selecting the supporting literature, which will be used to understand the current state of the art of the subject.

These procedures establish the repository where literature references would be sought, the protocols used to find publications, the criteria for inclusion or exclusion of papers, the procedures to evaluate the quality and compatibil-

ity of studies, and how the desired information would be extracted.

27.2.1 Research Questions

This work seeks to understand the relationship between population engagement and smart cities. From this, three questions were established to be answered along with this work:

Question 1: What are the main aspects that motivate and enable people to engage in Smart Cities participatory policies?

Question 2: What are the main difficulties in having a Smart City with actively participating citizens?

Question 3: How are these aspects that involve population engagement usually handled when designing a public participation policy in a Smart City?

27.2.2 Search Strategy

The selection of papers to compose the literature references on this work was carried out through two steps, where each one selected a set of publications. The first step was to search with a systematized protocol in a scientific repository. The publications selected in this step compose the *Protocol Search Set*. Subsequently, from the articles selected on this protocol, it was used the snowballing technique, which consists of searching for new papers from the bibliographic references of a publication set to establish a new group of articles, which will be called the *Snowballing Set*.

Searches were conducted through the Scopus Elsevier database. The search criteria presented was adjusted manually, until a satisfactory result was achieved. It searches in title, abstract, or keywords for both terms “public participation” and “smart city” accompanied for one of a set of words with meaning related to engagement.

The raw results obtained by the search passed through an inclusion and exclusion criteria filter, which is responsible for adding or discarding the articles in the Protocol Search Set, according to their compatibility with desirable results.

27.2.3 Inclusion and Exclusion Criteria

To filter the results obtained by the repository search aiming to select only the most relevant publications for this Systematic Literature Mapping, inclusion and exclusion criteria were established.

For a publication to be selected, it must meet at least one of the inclusion criteria (IC), and it must not meet any of the exclusion criteria (EC).

- **IC-01:** Address aspects related to the population’s engagement in public participation policies;
- **IC-02:** Address the relationship between governments and citizens in smart cities;
- **EC-01:** Publications that are not in English;
- **EC-02:** Publications that could not be accessed;
- **EC-03:** Publications before 2010;
- **EC-04:** Repeated publications;
- **EC-05:** Publications that are not academic articles;
- **EC-06:** Publications in journals of low relevance.

Journals and Conferences that did not have a percentile greater than or equal to 40 in the CiteScore Rank or did not have an H-5 index greater than or equal to 20 will be considered as of low relevance.

27.2.4 Snowballing

The selection of articles will be complemented by a *Snowballing Set*, where new papers will be selected from the bibliographic references of the publications chosen through the search protocol. In this step, it will be analyzed which publications were most cited among papers of the *Search Protocol Set*. A new selection set will be made with all publications that have been cited by at least 20% of the papers selected by the search protocol. In this case, the exclusion criteria related to the date, type, and relevance of the publication journal will be ignored, since the quality and relevance on the topic of these works are implicit as they are being cited by the publications selected following the Search Protocol.

27.2.5 Data Extraction

For each publication selected, both by the *Search Protocol Set* and the *Snowballing Set*, the title, authors, publication year, publication location, CiteScore rank percentile and H-5 index were extracted and tabulated.

In the body of the text of each publication, it was searched for items that answer the research questions.

27.3 Results

The search following the established protocol returned 24 results at Scopus database, which are listed in Table 27.1. Among these results, fourteen passed through the inclusion and exclusion criteria filter, composing the *Search Protocol Set*, and the other ten publications were discarded.

Table 27.2 presents the publications selected through the snowballing process. Nine publications were cited for at least three papers in the *Search Protocol Set*.

Table 27.1 The most recurring terms with relevance for this research between the two sets of publications were discussed individually, bringing to results the most relevant points found within the articles. Search protocol set

Refs.	Title	Relevance	Criteria	Status
[3]	Smart Citizens for Smart Cities – A User Engagement Protocol for Citizen Participation	percentile: 27th H5 Index: 32	IC-01, IC-02, and EC-02	Excluded
[4]	Rethinking public participation in the smart city	percentile: 74th H5 Index: 20	IC-01 and IC-02	Included
[5]	Towards more inclusive smart cities: Reconciling the divergent realities of data and discourse at the margins	percentile: 90th H5 Index: 32	IC-01 and IC-02	Included
[6]	Imaginarities of Sustainability: The Techno-Politics of Smart Cities	percentile: 96th H5 Index: 19	IC-02	Included
[7]	Access to ICT in Poland and the co-creation of Urban space in the process of modern social participation in a smart city-a case study	percentile: 80th	IC-01 and IC-02	Included
[8]	A review and reframing of participatory urban dashboards	percentile: 91st H5 Index: 22	IC-02	Included
[9]	Empowering Smart Cities Through Community Participation a Literature Review	percentile: 7th	IC-01, IC-02, EC-05, and EC-06	Excluded
[10]	Participation in e-government services and smart city programs: A case study of Malaysian local authority	percentile: 40th H5 Index: 8	IC-01 IC-02	Included
[11]	Social Media as Tool of SMART City Marketing: The Role of Social Media Users Regarding the Management of City Identity	–	IC-02, EC-05, and EC-06	Excluded
[12]	Social media as passive geo-participation in transportation planning–how effective are topic modeling & sentiment analysis in comparison with citizen surveys?	percentile: 90th H5 Index: 20	IC-02	Included
[13]	Local governance and access to urban services: Political and social inclusion in Indonesia	percentile: 16th	IC-01, IC-02, EC-05, and EC-06	Excluded
[14]	Towards a smart and sustainable city with the involvement of public participation-The case of Wroclaw	percentile: 80th	IC-01, IC-02	Included
[15]	The behaviours and job positions of citizens in smart cities’ development	percentile: 40th H5 Index: 8	IC-02	Included
[16]	Towards socially sustainable urban design: Analysing Acto R-area relations linking micro-morphology and micro-democracy	percentile: 53rd H5 Index: 13	IC-02	Included
[17]	Using social network data to improve planning and design of smart cities	percentile: 45th H5 Index: 14	IC-02	Included
[18]	Evolution of the smart city concept and of research into it [Ewolucja koncepcji i badania miasta inteligentnego]	percentile: 28th H5 Index: 9	IC-02 EC-01, EC-06	Excluded
[19]	Developing online illustrative and participatory tools for urban planning: towards open innovation and co-production through citizen engagement	H5 Index: 6	EC-06	Excluded
[20]	Information and communication technologies and public participation: interactive maps and value added for citizens	percentile: 99th H5 Index: 61	IC-02	Included
[21]	How are citizens involved in smart cities? Analysing citizen participation in Japanese “smart Communities”	percentile: 88th H5 Index: 19	IC-01, IC-02	Included
[22]	Virtual worlds as support tools for public engagement in urban design	percentile: 51th	IC-02	Included
[23]	A carbon footprint calculator for the municipal waste collection system of Bari	–	IC-01, IC-02 and EC-06	Excluded
[24]	Geoestrela: The next generation platform for reporting non-emergency issues -borough context	–	IC-02 and EC-06	Excluded
[25]	Smart governance, collaborative planning and planning support systems: A fruitful triangle?	percentile: 51th	IC-01, IC-02	Included
[26]	E-Society: A community engagement framework for construction projects	-	IC-02, EC-06	Excluded

27.3.1 Information and Communication Technology (ICT)

The term “Information and Communication Technology” (ICT) appears frequently in the analyzed publications, being cited by nine out of the fourteen publications of the *Search*

Protocol Set ([4–7, 14, 15, 20, 21, 25]) and by 6 out of the 9 publications of the *Snowballing Set* ([2, 27, 28, 30–32]).

The use of ICT to support the delivery of public service is highly associated with the ‘smart city’ definition [14], being responsible for providing the technological factor on a three-way interaction with public and institutions [15] and

Table 27.2 Snowballing set

Refs.	Title	Cited by
[2]	Smart Cities. Ranking of European medium-sized cities.	[10, 14, 15, 20, 21, 25]
[27]	Understanding smart cities: An integrative framework	[14, 15, 20, 21, 25]
[28]	Will the real smart city please stand up? Intelligent, progressive or entrepreneurial?	[5, 7, 20, 21, 25]
[1]	A Ladder Of Citizen Participation	[4, 8, 10, 16]
[29]	Beyond engagement and participation: User and community coproduction of public services	[10, 15, 21]
[30]	Conceptualizing smart city with dimensions of technology, people, and institutions	[10, 15, 20]
[31]	Programming environments: Environmentalty and citizen sensing in the smart city	[4, 6, 20]
[32]	Smartmentality: The Smart City as Disciplinary Strategy	[4, 5, 21]
[33]	Being a ‘citizen’ in the smart city: up and down the scaffold of smart citizen participation in Dublin, Ireland	[4, 5, 8]

even being considered a precondition for smart governance [25]. It has the potential to reduce participation costs [21] and provide new meaningful forms of participation for the public [4, 27], enhancing the role of the citizen in decision-making processes through the exchanging of information between population and public administration [7, 20, 21]. ICT is also broadly used to perform information sensing, process, and analyze data and provide administrative services to society [7].

27.3.2 Geographic Information System (GIS)

Geographic Information System (GIS) is a framework for gathering, managing, and analyzing geographic data [34]. They are designed to make spatial compositions of information and provide analysis and views of the data. It can be understood as a type of ICT, and it was often cited related to public participation in the publications analyzed.

Eight out of the fourteen articles ([5, 7, 8, 12, 14, 17, 20, 25]) in the *Search Protocol Set* discuss GIS benefits for Smart Cities or provide some practical example of application. On the other hand, the topic is not addressed in any of the 9 publications of the Snowballing Set. More specifically, Web applications based on maps that provide for population interaction to feed data are named Public Participation GIS (PPGIS) [7, 8, 14, 25] and are specially used for collect, from citizens, information based on location and also for offering easily understandable visualizations of data.

27.3.3 Open Data

Open data build trust between citizens and the public administration and promote accountability of the governing class [20]. This theme is raised by eight out of the fourteen papers of *Search Protocol Set* ([5–8, 10, 12, 14, 20, 21]) and by five out of nine papers of the *Snowballing Set* ([2, 27, 30, 32, 33]).

Public budget, safety, health, environment, economic, housing, education, fiscal and social indicators [8], actions and decisions of administration [33], services and transportation information [8] and even the data collected through public participation [6] are samples of relevant information that can be provided to the population. These data must be delivered in easily digestible formats to serve the population and government [8], and need to be appropriately elaborated, communicated, and used [20], covered by strict collection and processing standards to guarantee trust.

The main ways to make information available to the population are through digital platforms, such as websites and urban dashboards [6, 8, 10, 12, 20].

27.3.4 Institutional Structures and in-Person Meetings

Before the advent of ICTs, the most common methods of public participation were through face-to-face civic meetings [1], local referendums, public-private councils, and other institutional structures [14]. This process is described as exclusionary, as it fails to attract younger audiences and depends on people’s availability at scheduled times [16]. Technology has made popular participation easier and broader, making it possible for citizens to participate at any time and place [7, 21]. However, in-person assemblies are still used [4, 10, 12, 21, 33] and, in these cases, are more significant than technology in the decision-making processes [4], although they commonly have a low attendance and participation rate [10, 12, 16, 21].

27.3.5 Understanding Citizens

The development of smarter governance depends on deep knowledge about the citizens’ characteristics [15]. Urban planning must be a process where the city learns from itself, and from the people’s habits, needs, and desires [17].

Cultivation of aware and civic-minded citizens is strongly recommended as authorities’ priority [15]. Citizens should be engaged in decision-making processes in the early stages of a Smart City implementation [7] upstreaming the creation of ICT, to ensure participation before the development of tools that may follow certain planning pathways and limit public

power in decision-making processes [4]. The implementation of feedback channels is frequently cited by literature as a positive action [7, 8, 10, 14, 15, 20, 29, 31, 33].

27.3.6 Smart Cities and Corporate Interests

The participation of private companies in public governance is a recurring concern in the studies analyzed. Reference [28] raises that private interest in profit-making can overtake the public desire for citizens' inclusion, and [32] states that provision of technological infrastructures by private actors may cause separation between sealed-off technological enclaves and leftover marginalized spaces. Additionally, the control and processing of public big data by corporations can be driven to fulfill private interest [3].

27.4 Discussion

From the results obtained in the literature, the questions raised by the research were answered.

Question 1: What are the main aspects that motivate and enable people to engage in Smart Cities participatory policies?

The population's awareness about city governance and the importance of their participation in decision-making is the main factor in generating engagement, and awareness efforts must precede the application of technologies.

Among the tools used to improve engagement, several types of ICTs have already been developed, allowing the population to have a two-way information exchange channel with the government.

Question 2: What are the main difficulties in having a Smart City with actively participating citizens?

Ensuring active involvement by the population depends on accessibility to public participation tools. While traditional forms of consultation, through face-to-face meetings, have low attendance by citizens, ICTs can be a technological barrier for specific groups of the population. The government's lack of transparency to citizens is another factor that negatively affects people motivation.

Question 3: How are these aspects that involve population engagement usually handled when designing a public participation policy in a Smart City?

Typically, the smart city label comes with a wide technological investment in ICTs, accompanied by infrastructure installation and by use of information systems oriented to service provision. However, the investment not necessarily improves life quality for the population. When there is no planning with broad involvement of citizens from the early stages of the process, the decisions end up fulfilling private interests.

Not rarely, the government's effort to implement a smart city has more advertising than practical content, making it a form of tokenism, where the government invites the population to participate in governance without actually giving the necessary space in the decision-making process.

27.5 Conclusion

This Literature Systematic Mapping reinforced how the design of a Smart City is completely dependent on the activity sensing of its population. Without a deep understanding of citizens' wishes and needs, no effort to conduct public administration in a "Smart" way is valid. ICTs came as major promoters of recent advances in the topic, either ensuring popular participation or by creating models of administrative transparency and communication to the community. However, it is important to emphasize that popular participation must precede the use of technology itself, once the citizens must first discuss and approve how these new technologies will be used, given the direct impact it can bring on urban living.

References

1. S.R. Arnstein, A ladder of citizen participation. *J. Am. Plan. Assoc.* **35**, 216–224 (1969)
2. R. Giffinger, *Smart Cities Ranking of European Medium-Sized Cities* (Centre of Regional Science, Vienna University of Technology, Vienna, 2007)
3. B. Stelzle, A. Jannack, T. Holmer, et al., Smart citizens for smart cities – a user engagement protocol for citizen participation. *AISC* **1192**, 571–581 (2021)
4. A.M. Levenda, N. Keough, M. Rock, B. Miller, Rethinking public participation in the smart city. *Can. Geogr.* **64**, 344–358 (2020)
5. J.Y. Lee, O. Woods, L. Kong, Towards more inclusive smart cities: reconciling the divergent realities of data and discourse at the margins. *Geogr Compass* **14**(9), e12504 (2020)
6. T.R. Miller, Imaginaries of sustainability: the techno-politics of smart cities. *Sci. Cult.* **29**, 365–387 (2020)
7. P. Szarek-Iwaniuk, A. Senetra, Access to ICT in Poland and the co-creation of Urban space in the process of modern social participation in a smart city—a case study. *Sustainability* **12**(5), 2136 (2020)
8. O. Lock, T. Bednarz, S.Z. Leao, C. Pettit, A review and reframing of participatory urban dashboards. *City Cult. Soc.* **20**, 100294 (2020)
9. A. Kapoor, E. Singh, Empowering smart cities through community participation a literature review. *Lect. Notes Civil Eng.* **58**, 117–125 (2020)
10. L.S. Boon, J.A. Malek, M.Y. Hussain, Z. Tahir, Participation in e-government services and smart city programs: a case study of Malaysian local authority. *Plann. Malaysia* **18**, 300–312 (2020)
11. D. Petrikova, M. Jaššo, M. Hajduk, Social media as tool of SMART city marketing, in *Smart Governance for Cities: Perspectives and Experiences*, (Springer, Cham, 2020), pp. 55–72
12. O. Lock, C. Pettit, Social media as passive geo-participation in transportation planning—how effective are topic modeling & sentiment analysis in comparison with citizen surveys? *Geo-Spatial Inf. Sci.* (2020). <https://doi.org/10.1080/10095020.2020.1815596>

13. W. Salim, M. Drenth, Local governance and access to urban services: political and social inclusion in Indonesia. *Adv. 21st Century Hum. Settlements* 153–183 (2020)
14. D. Bednarska-Olejniczak, J. Olejniczak, L. Svobodová, Towards a smart and sustainable city with the involvement of public participation – the case of Wrocław. *Sustainability* **11**(332) (2019)
15. S. Lim, J.A. Malek, M.Y. Hussain, Z. Tahir, The behaviours and job positions of citizens in smart cities' development. *Plann. Malaysia* **17**, 133–145 (2019)
16. R. Timmerman, S. Marshall, Y. Zhang, Towards socially sustainable urban design: Analysing Actor-Area relations linking micro-morphology and micro-democracy. *Int. J. Sustain. Dev. Plan.* **14**, 20–30 (2019)
17. R. Pérez-Delhoyo, H. Mora, J.F. Paredes, Using social network data to improve planning and design of smart cities, in *WIT Transactions on the Built Environment*, (WIT Press, Department of Building Sciences and Urbanism, University of Alicante, Spain, 2018), pp. 171–178
18. G. Masik, D. Studzińska, Evolution of the smart city concept and of research into it. *Prz. Geogr.* **90**, 557–571 (2018)
19. V. Oksman, M. Kulju, Developing online illustrative and participatory tools for urban planning: towards open innovation and co-production through citizen engagement. *Int. J. Serv. Technol. Manag.* **23**, 445–464 (2017)
20. D. Gagliardi, L. Schina, M.L. Sarcinella, et al., Information and communication technologies and public participation: interactive maps and value added for citizens. *Gov. Inf. Q.* **34**, 153–166 (2017)
21. B. Granier, H. Kudo, How are citizens involved in smart cities? Analysing citizen participation in Japanese “smart Communities”. *Inf. Polity* **21**, 61–76 (2016)
22. A. Jutraz, T. Zupancic, Virtual worlds as support tools for public engagement in urban design. In: R. G., J. F., J. S., S. G. (eds). Kluwer Academic Publishers, University of Ljubljana, Zoisova 12, Ljubljana, 1000, Slovenia, pp. 391–408 (2015)
23. S. Digiesi, G. Mossa, G. Mummolo, R. Verriello, A carbon footprint calculator for the municipal waste collection system of Bari. *AIDI – Italian Association of Industrial Operations Professors, Department of Mechanics, Mathematics and Management Engineering, Polytechnic University of Bari, viale Japigia, 182–60100, Bari, 70126, Italy*, pp. 124–129 (2015)
24. L.P.A.C.N. Parreira, T.M.P. Santos, S.P. Da Silva Vendeirinho, Geoestrela: The next generation platform for reporting non-emergency issues -borough context. In: P. I., L. R., H.F. B., et al (eds). IADIS, Junta de Freguesia da Estrela, Lisbon, Portugal, pp. 206–210 (2015)
25. Y. Lin, S. Geertman, Smart governance, collaborative planning and planning support systems: A fruitful triangle? In: R. G., J. F., J. S., S. G. (eds). Kluwer Academic Publishers, Department of Human Geography and Planning, Utrecht University, Utrecht, Netherlands, pp. 261–277 (2015)
26. S. Kinawy, T.E. El-Diraby, E-Society: A community engagement framework for construction projects. Department of Civil Engineering, University of Toronto, 35 St. George St., GB105, Toronto, ON M5S1A4, Canada, pp. 676–686 (2010)
27. H. Chourabi, T. Nam, S. Walker, et al, Understanding smart cities: An integrative framework. In: *Proceedings of the Annual Hawaii International Conference on System Sciences* (2012)
28. R.G. Hollands, Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? *City* **12**(3), 303–320 (2008)
29. T. Bovaird, Beyond engagement and participation: user and community coproduction of public services. *Public Adm. Rev.* **67**, 846–860 (2007)
30. T. Nam, T.A. Pardo, Conceptualizing smart city with dimensions of technology, people, and institutions, in *ACM International Conference Proceeding Series*, (2011), pp. 282–291
31. J. Gabrys, Programming environments: environmental and citizen sensing in the smart city. *Environ. Plan. D Soc. Space* **32**, 30–48 (2014)
32. A. Vanolo, Smartmaturity: the smart city as disciplinary strategy. *Urban Stud.* **51**, 883–898 (2014)
33. P. Cardullo, R. Kitchin, Being a ‘citizen’ in the smart city: up and down the scaffold of smart citizen participation in Dublin, Ireland. *GeoJournal* **84**(1), 1–13 (2019)
34. ESRI, What is GIS? <https://www.esri.com/en-us/what-is-gis/overview>. Accessed 28 Oct 2020

Use of Crowdsourcing Questionnaires to Validate the Requirements of an Application for Pet Management

28

Vitor S. Vidal, Marco Aurélio M. Suriani, Rodrigo A. S. Braga, Ana Carolina O. Santos, Otávio S. Silva, and Roger J. Campos

Abstract

Mobile applications usually fail to deliver the right set of features for its users, either by not offering necessary functionalities or by offering unnecessary ones. In the last few years, some modern Requirements Engineering methods have been created to better design mobile applications. One of these methodologies is the Crowdsourcing Requirements Engineering, based on short cycles of implementations and feedbacks by a large group of actual users. This work aims to validate the requirements of a mobile application for the management of domestic animals, through the use of crowdsourcing questionnaires. Two questionnaires were designed and implemented to assess the needs of the users. The first questionnaire verified functionalities provided by similar softwares, while the second one verified requirements established by the authors.

From the analysis of the data gathered, all requirements, except for one, were validated. Finally, after the final functionalities of the software were defined, an alpha version of the software could be created.

Keywords

Software engineering · Mobile applications · Requirements engineering · Crowdsourcing ·

Questionnaires · Likert scale · Functional requirement · Requirement validation · Pet management · Research method

28.1 Introduction

The specification of requirements for software systems, according to the IEEE guidelines [1], is fundamental for the development of correct software that have the specific functionalities users need. The publication establishes an international standard for specifying requirements in the area of software engineering, helping to produce consistent results. The IEEE guidelines also state that failing to understand customer needs is one of the most common problems in the requirements specification process.

The book [2] portrays some of the problems of specifying requirements for software as a product, which are the production of a software that cannot communicate directly with the target audience and the non-validation of the established requirements. The usual approach to these problems are, respectively, to determine a specific user experience, as exemplified on [3], and to adopt a method to validate the requirements.

However, the rising of the mobile applications has significantly changed the techniques of Software Development in the last few years. According to [4], the traditional methodologies of Software Development include a set of sequential stages in a predefined order, while the modern methodologies, or Agile Development, include a set of non-sequential and iterative steps to promote flexibility, adaptability and efficiency. A comparison present in [5] states that traditional Software Engineering focuses on deciding all the requirements before coding and does not welcome changing them during the projects progress, while Agile Development focuses on short deliveries and encourages the change of

V. S. Vidal · M. A. M. Suriani · R. A. S. Braga (✉) · O. S. Silva
R. J. Campos
Institute of Science and Technology, Federal University of Itajuba,
Itabira, MG, Brazil
e-mail: vtorvidal@unifei.edu.br; marcosuriani@unifei.edu.br;
rodrigobraga@unifei.edu.br; otaviosoressilva@unifei.edu.br;
rogercampos@unifei.edu.br

A. C. O. Santos
Integrated Engineering Institute, Federal University of Itajuba, Itabira,
MG, Brazil
e-mail: anasantos@unifei.edu.br

requirements. Besides that, Agile Development involves customers and users continuously, with shorter cycles of releases and feedback.

According to [6], the Requirements Engineering (RE) has also switched from a traditional approach to a modern one: Crowd-based requirements engineering (CrowdRE). The authors argue that traditional RE methods fail to reach large and heterogeneous groups of users and to consider some valuable resources for RE. The CrowdRE, on the other hand, aims to motivate a large crowd of users and other stakeholders to continuously use the application and provide feedback, resulting in some iterations of development seeking to identify the customer value [7].

Thus, crowdsourcing is a process of obtaining information by asking for contributions from a group of people, usually online and through the use of questionnaires. According to [8], the LEGO company managed to evolve the product development process with the help of its customers using crowdsourcing. Results like this show that it can be feasible to apply questionnaires using crowdsourcing to validate conceptual products, such as the requirements of an application. Besides that, it is observed in [9] that the application of questionnaires is quite effective in scenarios where personal interaction is either not possible or infeasible.

The importance of requirements validation in the application's market success is evident, since it verifies which functional requirements will be really useful for the interested parties. This study shows the use of a crowdsourcing methodology to validate the requirements of a mobile software for the management of domestic animals. This validation is employed by the application of two questionnaires, in order to verify if the proposed functionalities are adequate to facilitate the management of the interviewees' pets. Thus, the objective of this work is to contribute to the area of research in Software Engineering, by showing the validation of requirements of the proposed software through crowdsourcing and the application of questionnaires. The development of the alpha version of the proposed software based on the validated requirements of this work is present in [10].

This paper was organized as follows. Section 2 shows related work. The Sect. 3 shows the questionnaires and explains the methodology adopted for their elaboration. Section 4 shows the data gathered and the final definition of requirements. Lastly, Sect. 5 shows the conclusion reached with this study.

28.2 Related Work

28.2.1 Requirements Engineering Processes

The work [11] is the foundation used by the international standard for Requirements Engineering. It defines a require-

ment as a statement expressing a need and its associated restrictions and conditions. The requirement must be described specifying the implicit functionality restrictions and which conditions allow the elaboration of a system functionality.

Therefore, it is essential that the requirements are specifically and unequivocally necessary, ensuring a clear explanation of the functionality of each item present in the software. As can be seen in [9], a failure in the definition of requirements can result in low quality software and, thus, in user discontent. The inclusion of user feedback in the requirements collection stage benefits the process of application development.

The paper [9] also portrays the difficulties of Requirements Engineering in the field of mobile applications. The study sought to validate methodologies that can be applied to the collection of requirements for mobile applications through a search in previous publications. This work shows that the Requirements Engineering must be worked with caution when it comes to mobile applications, since the users become dissatisfied with applications lacking the necessary functionalities. Such dissatisfaction may result in a decrease in the number of users of a given app.

The publication [2] emphasizes the problems caused by the failure to apply the Requirements Engineering, such as the development of systems without critical properties and the presence of unsolicited functionalities. In addition, it is formalized that any requirement related to the behavior of a system functionality is a functional requirement. According to the publication, it is essential to understand the development context of the system. Finally, the description of the developed system should be made explicit, making clear its purpose and scope and explaining the context in which the software will be used.

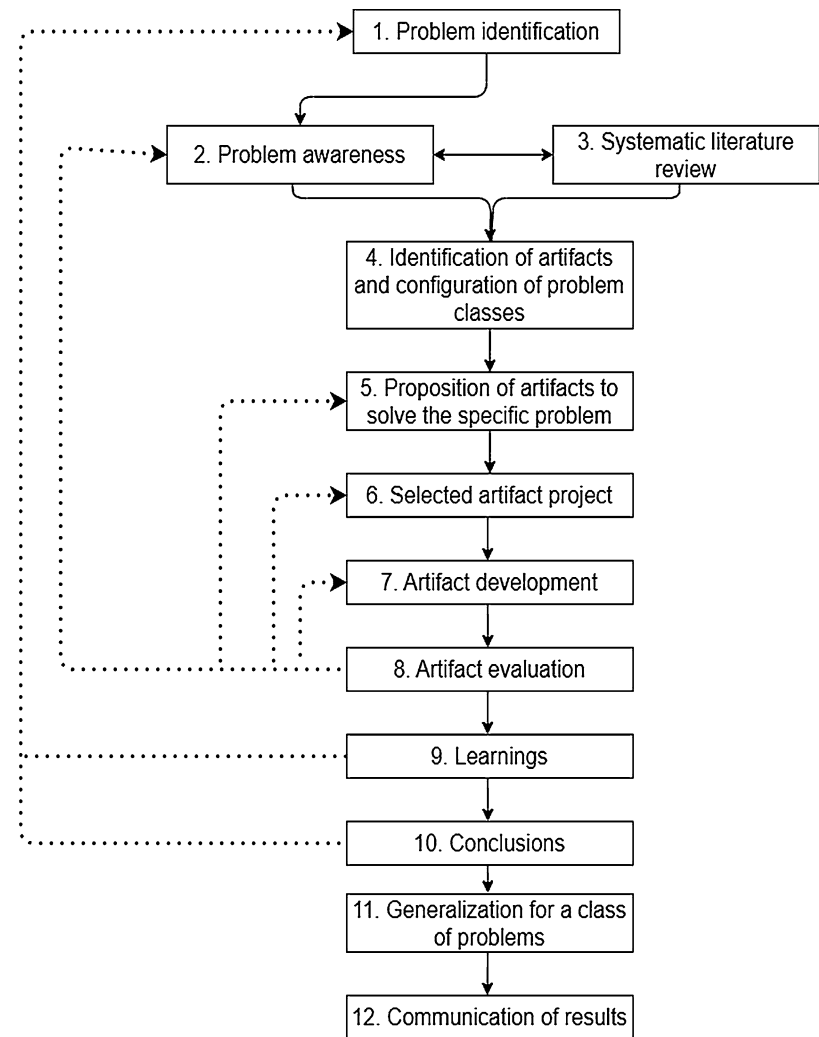
28.2.2 Elaboration of Questionnaires

The research conducted in [3] concludes that mobile applications should not have optional features. Only the necessary functionalities should be in the final software, facilitating the user experience in the developed interface. Therefore, validating the requirements of an application by using some method has become a good practice in Software Engineering.

The use of crowdsourcing in Software Engineering as a method to validate requirements is increasingly common, according to [12]. This is due to the growth of online communication platforms, which allow data collection from a more diverse and comprehensive audience, resulting in a validated product that is capable of meeting the demands of different types of users.

The study [13] grounds the concepts of questionnaire application and crowdsourcing for the requirements validation. It states that the main objective of a questionnaire is to

Fig. 28.1 Steps of the research method used. (Adapted from [14])



assess the opinions and interests of the interviewees and, therefore, it is appropriate for the validation of a project's requirements.

The formulation of the questions must give to the interviewees a wider range of possible answers. The Likert Scale, also discussed in [13], defines the possible answers to a question through an interval that varies between the two possible extremes for response data. This scale allows more comprehensive responses, and it validates users' interests in a more useful way.

28.3 Development

The research method selected to conduct this work is an adaptation of the model proposed by [14], shown in the Fig. 28.1. This model has the steps common to various research methods, namely:

- Problem definition
- Proposal of suggestions to solve the problem
- Development of the artifact
- Evolution of the artifact.

In Fig. 28.1, steps 1–3 define the requirements validation problem and the possible solutions that exist for it, which are proven in an academic way. Step 4 corresponds to the selection of which of these possible solutions can be adapted to the problem in question, such as the application of two questionnaires using crowdsourcing. Steps 5–7 describe the elaboration of the questionnaire questions according to the initial proposed requirements and the collection of this data. Step 8 represents the data analysis of the responses obtained and the definition of the final requirements. The remaining steps show the lessons learned from this study, as well as a generalization of this methodology to similar problems and the main results obtained.

28.3.1 Proposal of Functionalities and Requirements

The software [15–17] were used as examples to define the proposed initial functionalities. Based on them, we defined the following list of initial features:

- Set daily reminders: Monitor your pet’s activity through a calendar
- Control of vaccinations
- Set notifications
- Available as an mobile app
- Register each animal separately

We also established some viable functional requirements to be implemented, considering the scope of the current work. All requirements considered viable can be seen in the following list. It is noted that some of these features were considered essential for the development of the application and, therefore, were not included in the applied questionnaires:

- Organize animals in multiple profiles, to separate the data for each animal and to decrease the number of information present in the application interface at any given time.
- Allow the creation of personalized notifications, specifying dates, times and repetition patterns for such notifications. This system allows the creation of notifications of times to feed the animal, times to walk the animal, among other possible notifications.
- Allow the creation of written reminders, which are shown prominently in the application interface.
- Create a list of the most common vaccines that animals of a particular species should have. Through this list, it will be possible to inform the user when their pet should have a common vaccine. It is necessary to collect the animal’s birth date to implement this functionality.
- Allow the creation of an account for each user. The account can be created through a user’s email or through integration with a user’s Google account.
- Create a form to register a new animal (essential).
- Make frequent automatic backups of each user’s account (essential).

28.3.2 Elaboration of the First Questionnaire

This work also establishes the use of crowdsourcing methodologies, which correspond to the involvement of a group to accomplish a task or achieve a goal. Through crowdsourcing, we used questionnaires to provide analytical evidence about whether the requirements established for a pet management application are necessary for interested parties or not. It is considered an interested party to the project users who have

Table 28.1 Questions applied in the first questionnaire

Question	Response type
How many pets do you have?	Numeric
If you have more than one animal, would you like the information for each animal to be shown separately?	Binary
Would you like to receive notifications of times to feed your pet?	Binary
Would you like to receive notifications of the days when your pet should bathe?	Binary
Would you like to write down reminders while using the app?	Binary
Would you like to receive notifications of the dates your pet should be vaccinated?	Binary
Would you like to set reminders using a calendar?	Binary

a smartphone with the Android operating system, version 6.0 or higher, and who are interested in using an application to help in the management of their pets.

A functionality is not considered validated if there are more negative than positive responses for its implementation. From the analysis of the responses obtained with the application of the questionnaires, the functionalities that will be present are defined.

Two questionnaires were applied to validate the established requirements. The first questionnaire aimed to prove the existence of a target audience interested in the proposed application and to verify if the functionalities provided by similar software were useful for them. The questions asked in the first questionnaire were organized according to Table 28.1. The methodology adopted in this questionnaire utilized only binary responses to validate the functionalities of the application.

28.3.3 Elaboration of the Second Questionnaire

The second questionnaire, shown in Table 28.2, was created to obtain more comprehensive answers, and to better represent the needs of users. In order to obtain more detailed answers, the Likert Scale was defined as the type of answer, with a scale graduated at five levels, which offered more options for the interviewees.

28.4 Results

The total of 15 and 20 people answered the first and the second questionnaire, respectively. The average number of pets per user was 2.64 and 2.75 in each questionnaire. Table 28.3 shows the answers obtained in each question in the first

Table 28.2 Questions applied in the second questionnaire

Question	Response type
How many pets do you have?	Numeric
How important is it for you that the information for each of your animals is shown separately?	Likert Scale
How often would you like to receive notifications of times to feed your pet?	Likert Scale
How often would you like to receive notifications of the days when your pet should bathe?	Likert Scale
How often would you like to write down reminders while using the app?	Likert Scale
How often would you like to receive notifications of the dates your pet should be vaccinated?	Likert Scale
How often would you like to set reminders using a calendar?	Likert Scale
How often would you like to receive alerts to buy feed?	Likert Scale
Would you like to fill in the details of the animal in detail, adding information such as weight and birth data?	Likert Scale

Table 28.3 Analysis of responses obtained in the first questionnaire

Question	Negative answers	Positive answers
If you have more than one animal, would you like the information for each animal to be shown separately?	20%	80%
Would you like to receive notifications of times to feed your pet?	13.3%	86.7%
Would you like to receive notifications of the days when your pet should bathe?	13.3%	86.7%
Would you like to write down reminders while using the app?	20%	80%
Would you like to receive notifications of the dates your pet should be vaccinated?	0%	100%
Would you like to set reminders using a calendar?	13.3%	86.7%

questionnaire. All the questions had 80% or more of positive answers.

Table 28.4 refers to the data analysis of the second questionnaire answers. Its data shows that the requirement to add the reminder functionality is not valid for the proposed application and, therefore, should not be implemented.

28.5 Conclusion

We have proposed two questionnaires to validate the requirements for a mobile application for pet management using a Crowdsourcing Requirements Engineering methodology.

Table 28.4 Analysis of responses obtained in the second questionnaire

Question	Negative answers	Impartial answers	Positive answers	Requirement
How important is it for you that the information for each of your animals is shown separately?	15%	10%	75%	Valid
How often would you like to receive notifications of times to feed your pet?	30%	20%	50%	Valid
How often would you like to receive notifications of the days when your pet should bathe?	25%	10%	65%	Valid
How often would you like to write down reminders while using the app?	40%	35%	25%	Invalid
How often would you like to receive notifications of the dates your pet should be vaccinated?	5%	15%	80%	Valid
How often would you like to set reminders using a calendar?	10%	50%	40%	Valid
How often would you like to receive alerts to buy feed?	10%	20%	70%	Valid
Would you like to fill in the details of the animal in detail, adding information such as weight and birth data?	30%	10%	60%	Valid

The first questionnaire aimed to verify if the functionalities provided by similar software were useful for the users, while the second one aimed to obtain information about some functional requirements established in an earlier development phase. We applied both questionnaires to some users and we analyzed their answers.

The questionnaires have successfully assessed the needs of the users, despite being applied to a small number of people. All requirements present in the questionnaires were validated, except for one functionality related to writing down reminders while using the app.

Based on those functionalities, the alpha version of a mobile pet management application could be created, as shown in [10]. The development of this alpha version involved the elaboration of the entity-relationship model, the elaboration of the rapid prototyping of the application interfaces and the elaboration of the activity diagrams representing the user interaction.

Future studies will consist in the development of the final version of the proposed application, emphasizing the topics of design and accessibility of mobile applications.

References

1. I.C. Society, IEEE recommended practice for software requirements specifications. IEEE Access (1998)
2. K. Pohl, C. Rupp, Requirements engineering fundamentals: a study guide for the certified professional for requirements engineering exam. IREB Compliant (2015)
3. K. Kuusinen, T. Mikkonen, On designing ux for mobile enterprise apps, in *Euromicro Conference on Software Engineering and Advanced Applications*, vol. 1, (IEEE, 2014), pp. 1–8
4. L.K. Shinde, Y.S. Tangde, R.P. Kulkarni, Traditional vs. modern software engineering – an overview of similarities and differences. *Adv. Comput. Res.* **7**, 187–190 (2015)
5. A. Aitken, V. Ilango, A comparative analysis of traditional software engineering and agile software development, in *Hawaii International Conference on System Sciences*, (IEEE, 2013), pp. 4751–4760
6. E.C. Groen, N. Seyff, R. Ali, F. Dalpiaz, J. Doerr, E. Guzman, M. Hosseini, J. Marco, M. Oriol, A. Perini, M. Stade, The crowd in requirements engineering – the landscape and challenges. *IEEE Softw.* **34**, 34–52 (2017)
7. A.C.O. Santos, C.E.S. da Silva, R.A.D.S. Braga, J.É. Corrêa, F.A. de Almeida, Customer value in lean product development: conceptual model for incremental innovations. *Syst. Eng.* **23**(3), 281–293 (2020)
8. J. Wei, S. Liu, et al., The customer dominated innovation process: involving customers as designers and decision-makers in developing new product. *Des. J.* **22**(3), 299–324 (2019)
9. H. Dar, M.I. Lali, H. Ashraf, M. Ramzan, T. Amjad, B. Shahzad, A systematic study on software requirements elicitation techniques and its challenges in mobile application development. *IEEE Access* **6**, 859–863 (2018)
10. V. S. Vidal, Uso de questionários para validar os requisitos de um aplicativo para gestão de animais de estimação, Technical Report. Federal University of Itajuba (2020)
11. I. O. for Standardization ISO, ISO/IEC/IEEE 29148: 2018-systems and software engineering – life cycle processes – requirements engineering (2018)
12. K. Mao, L. Capra, M. Harman, Y. Jia, A survey of the use of crowdsourcing in software engineering. *J. Syst. Softw.* **126**, 57–84 (2016)
13. T. Nemoto, D. Beglar, Likert-scale questionnaires, in *JALT 2013 Conference Proceedings*, (2014), pp. 1–8
14. A. Dresch, D.P. Lacerda, J.A.V.A. Júnior, *Design science research: método de pesquisa para avanço da ciência e tecnologia* (Bookman Editora, 2015)
15. Animal care – 11pets, 11pets (2020) [Online]. Available: <https://www.11pets.com>
16. Kennel software with free trial, Revelation Pets (2020) [Online]. Available: <http://revelationpets.com/>
17. Mytrackpet – animal control management system, MyTrackPet (2020) [Online]. Available: <https://mytrackpet.com/>

Discovery of Real World Context Event Patterns for Smartphone Devices Using Conditional Random Fields

29

Shraddha Piparia, Md Khorrom Khan, and Renée Bryce

Abstract

Mobile applications are Event Driven Systems that react to user events and context events (e.g. changes in network connectivity, battery level, etc.) The large number of context events complicate the testing process. Context events may modify several context variables (e.g. screen orientation, connectivity status, etc.) that affect the behavior of an application. This work examines a data set of real-world context changes on Android phones. We collect every context event that occurs on the mobile devices of 58 Android users over 30 days to identify complex relationships and patterns. This work uses Machine Learning (ML) techniques including Conditional Random Fields (CRFs) and Deep Neural Networks (DNNs) to predict sequence labels for context events. These techniques are compared to Majority Baseline (MB). The trade-offs among these methods reveal that CRF is the most effective technique for sequence prediction/labeling of the data-set. Future work may apply the data collection strategy and ML techniques to domains for emerging technologies in areas such as Internet of Thing, smartwatches, and autonomous vehicles.

Keywords

Android applications · Conditional random fields · Context-aware applications · Deep neural networks · Machine learning · Neural network architectures · Sequence prediction · Software testing

S. Piparia (✉) · Md. K. Khan · R. Bryce
Computer Science and Engineering, University of North Texas,
Denton, TX, USA
e-mail: ShraddhaPiparia@my.unt.edu; MdKhorromKhan@my.unt.edu;
Renee.Bryce@unt.edu

29.1 Introduction

Security and testing concerns have grown with the widespread use of smartphone applications. These applications are context aware and change their behaviour in response to information provided by context input sources (e.g. WiFi, GPS, Bluetooth, screen orientation, etc.) Android and iOS are two leading smartphone operating systems with Google Play dominating the market [18]. The Google Play store has seen a rapid increase in number of downloads indicating how vast Android has grown, thereby increasing the requirement for an efficient testing method. Due to advancements in the hardware and software industry, smartphone devices continue to have increasingly complex operating systems, faster processors, always-on connectivity, and a diverse array of on-board hardware sensors and peripherals. These rapid advancements increase the complexity of smartphone apps and the processes required to develop and test them.

Android devices are capable of generating 144 context events for apps to respond [1]. These context events affect the functioning of the application not only at the launch of the application but throughout their life cycle. For instance, consider a version of the Android Wikipedia app that crashes when an user tries to save a page without an internet connection [9]. Organizations need to test for incorrect application behavior in response to contextual changes.

Cost effective testing of context-aware smartphone applications is challenging because it is difficult to identify specific sequences of context events that will negatively affect an application's behavior. In this work, we seek to identify context event sequences that occur in the real world and incorporate these sequences into automated testing methodologies. Context event sequences based on real-world smart-

phone context data can provide insight into the likelihood that particular context event sequences will occur during an application's execution. The prediction of real-world context event sequences can help ensure that the most likely context event sequences are tested respectively and relates to users' perceived reliability of apps.

The contributions of this paper are (1) examination of a real-world context event data-set from Android users and (2) application and empirical study of ML techniques to construct context event sequences from real data. To collect data, we implemented an Android application, ContextMon, to listen and record context events from each volunteer's smartphone device. The smartphone app listens for context events on its host device and sends details about these events to a remote server. This app uses event broadcast mechanisms inherent in the Android smartphone operating system. ContextMon collected context event data from 58 Android devices over 30 days. We then apply (1) CRF, a discriminative model, to model the conditional probability combines the advantages of classification and graphical modeling, (2) LSTM, a recurrent neural network model, to remember complex dependencies from past context events, (3) GRU, a gated recurrent neural network model, to capture features just like LSTM but without a memory unit, and (4) Majority Baseline, a simple model to estimate baseline values.

In the remainder of the paper, Sect. 29.2 provides background on context aware applications and sequence prediction/labeling of context events using MB, RNNs, and CRF techniques. Section 29.3 describes how we set up the infrastructure for data collection and analysis. Section 29.4 discusses our research questions and methodology. Section 29.5 analyzes the prediction success of the techniques. Section 29.5.1 summarizes the results and highlights how future work may extend our techniques to other context aware domains such as IoT, smart watches, and autonomous vehicles.

29.2 Background and Related Work

This section discusses context aware systems, hidden markov model, conditional random fields, and neural network architectures.

29.2.1 Context Aware Applications

A context event is defined as a 2-tuple (c, a) where c is a context category and a is a context action. Table 29.1 shows a broad categorization of context. A context event may affect the manner in which the smartphone application responds immediately or to subsequent user interaction events. For example, when a user uses an app that requires internet access, a network disconnection may immediately impede the progress of a task. An app may change performance due

Table 29.1 Categories of context event

Context event type	Example
Device hardware events	Changes in battery power levels, Changes in network connectivity, etc.
Operating system events	System reboot, Screen orientation changes, etc.
Device sensor events	Location changes(GPS), Humidity changes(hygrometer).
Typical smartphone events	Arrival of a phone call or a SMS message.
Application events	Arrival of an email or smartphone notifications.

to a context change such as reducing the rate at which it sends data over a network when battery levels go below a certain threshold. An app may respond to a context change in a delayed manner such as an app that works offline for some tasks and then requires internet for others. In the Android platform, this sort of behavior is implemented as a BroadcastReceiver to listen for such context events.

29.2.2 HMM and CRF

Hidden Markov Models (HMMs) [15] are probabilistic sequence models with widespread use in natural language processing tasks such as part of speech tagging, sequence prediction, information extraction and speech recognition [6]. HMMs are appealing because it is relatively more straightforward for machines to learn its parameters and for humans to interpret. However, the drawback of HMM is that it is difficult to model arbitrary, dependent features of the input sequence. HMMs assume that the features are generated independently by a hidden process. To overcome this drawback, we propose use of conditional random field (CRF) [19]. CRFs are sequence models which are discriminative in nature as opposed to generatively trained HMMs. CRFs maximize the conditional likelihood whereas HMM maximize the joint likelihood thereby optimizing testing accuracy.

29.2.3 Deep Neural Networks

Recurrent neural networks (RNNs) are networks specifically designed to recognize patterns in sequences of data such as text, spoken word, genomes, etc. The two types of RNNs used in this study are Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). Recent data is often sufficient to make correct predictions, but sometimes past data is needed. LSTMs are RNNs which are capable of learning long-term dependencies. LSTM has a chain-like neural network layer with input, output, and forget gates. GRU is a gating mechanism in RNNs with update and reset gates. They control the flow of information like LSTM, but

without having to use a memory unit. GRUs are simpler and often considered to be more efficient than LSTMs.

29.2.4 Related Work

Prior research proposes the use of ML algorithms for software testing. Onan [13] evaluated web page classification accuracy of four different algorithms (Naive Bayes, K-nearest neighbor algorithm, C4.5 algorithm, and FURIA algorithm) using different feature selection techniques and ensemble learning methods for web page classification. Gogar et al. [8] introduced a site-independent web wrapper which uses CNNs to extract information. The authors proposed a spatial-text method to combine both text and visual data into one neural network.

While approaches proposed by Onan and Gogar et al. were designed specifically for web applications, GUI testing using AI has been studied for smartphone applications as well. Memon [11] automated GUI regression testing using AI planning. GUI test cases are represented as a pair of initial and expected states including the sequences of states. An AI driven software testing approach is presented by Arbon [3] where the application state of smartphone applications are identified, test inputs are applied, and behavior of the application verified. Three ML approaches used for this work were random forests, decision tree learning, and ANNs. Dionny et al. [16] created a test flow generation system which is trainable using AI and ML techniques. A language and grammar define test cases as an abstract sequence. Arbon et al. [2] present an Abstract Intent Test (AIT) language to manually define test cases using domain concepts. The elements of AIT in this approach are similar to the one proposed by Dionny et al.

RNNs are widely used approach for sequence prediction in various domains. One such work by Graves [10] demonstrates the ability of RNNs with LSTM units to generate complex sequences with long range structure. The network performs just as well when predicting one word at a time than predicting one character at a time. Gers et al. [7] demonstrate that LSTM outperforms traditional RNNs and are capable of learning context-free and context sensitive languages. While these techniques make use of ML in software testing, the context-sensitive nature of smartphone applications was not taken into account. In this work, we use RNNs for sequence prediction of real context event sequences.

29.3 Research Methodology

The research methodology is comprised of three main activities—the data collection process, data pre-processing and context modeling. Each of these activities are described in detail as follows:

29.3.1 Data Collection Process

The data collection process consists of an Android application which listens for context events on a host smartphone device, an on-device database that stores information about detected context events, and a remote database that stores context events aggregated from multiple devices.

- *Android smartphone application.* An application runs in the background and collects context events from host devices. It uses the broadcast receiver to listen for event broadcasts sent by the operating system. Information regarding each broadcast is contained in its associated intent.
- *SQLite.* SQLite [17] is a software library that implements a self-contained, server-less, zero-configuration, transactional SQL database engine. This database engine is part of the Android OS. Context events are stored locally on the smartphone device using this database engine.
- *Remote MySQL database.* A MySQL [12] database receives and aggregates all context events gathered from the devices. Context events are retrieved and removed from the local SQLite database on participants' smartphone devices every 15 minutes and sent to a remote database server. An HTTP endpoint was implemented using the PHP programming language. This endpoint is receives and prepares the data to load into the MySQL database.

29.3.2 Data Preprocessing

We retrieve the data from the remote database and separate it by user. The user data is processed and represented in a human readable format. This process includes two stages:

Stage 1: Data Transformation and Representation We extract context data with timestamps. As shown in Table 29.2, an Android broadcast contains an *action* and *extra*. The *action* field indicates the nature of the context event, while the *extra* field provides additional information about the event. The information contained in the *extra* field is not represented in a manner that is easy to analyze. To make the data processing easier, raw data for each context event was transformed into a representation that is more amenable to analysis. This transformed representation is defined in Definition 29.1. Context event data for each smartphone device is represented as a sequence of 2-tuples. Table 29.3 shows an example of such a sequence using the transformed representation. Context events that are not of interest to our analysis are filtered out.

Definition 29.1 A context event is a 2-tuple (c, a) where c is a context category and a is a context action.

Table 29.2 Data collected for each context event

Data label	Sample value	Meaning
device_id	ANONYMISED ID	A unique identifier for a particular smartphone device. This unique ID is retrieved from each device.
action	android.intent.action.HEADSET_PLUG	Context event indicating that a wired headset has been plugged/unplugged into/from the device.
extras	name: h2w state: 1 microphone: 1	Additional information about the context event indicating: (1) whether the headset was plugged or unplugged (2) the type of headset (3) whether or not the headset has a microphone.
timestamp	2015-12-04 07:10:27	The time and date when this context event occurred.
android_version	4.4.2	The Android OS version on the device from which this context event was recorded.

Table 29.3 Transformed context event representation

Representation	Meaning
('data_connection', 'lte_connected')	Smartphone device is connected to LTE network
('time_date', 'time_set')	System time was set (changed)
('wifi_device', 'on')	Wifi device is on

Table 29.4 Difference between sequence with redundant events and sequence without redundant events (redundant events in bold)

Sequence with redundant events	Sequence without redundant events
('data_connection', 'lte_connected') ('data_connection', 'lte_connected') ('time_date', 'time_set') ('wifi_device', 'on') ('audio', 'audio_effects_closed') ('audio', 'audio_effects_closed')	('data_connection', 'lte_connected') ('time_date', 'time_set') ('wifi_device', 'on') ('audio', 'audio_effects_closed')

Stage2: Remove Redundant Context Events A context event in a sequence is redundant if it has appeared more than once in consecutive order. Redundant context events were eliminated from the sequence to avoid self loops. Table 29.4 demonstrates the difference between context event sequence with and without redundant events.

29.3.3 Context Modeling

Sequence prediction uses previous events to predict the next most likely events. Sequence prediction is different than the other classification and regression problems because it requires learning dependencies that exist between the order of events. We model context events using CRF, LSTM, and GRU. NNs and CRFs are discriminatively trained probabilistic models. While neural networks are best known for tasks involving classification, they are also used to predict sequences.

29.4 Empirical Study

The experiments analyze performance of MB, CRF, LSTM, and GRU.

29.4.1 Research Questions

Our research questions examine:

RQ1: How does majority baseline perform in terms of precision, recall, and F-1 score for sequence prediction of real-world context events?

RQ2: Does a neural network increase precision, recall, and F-1 score compared to majority baseline?

RQ3: Does a CRF increase precision, recall, and F-1 score compared to majority baseline?

29.4.2 Study Participants and Data Description

Context event data was collected in an uncontrolled environment during the course of a smartphone device's regular everyday use. Table 29.2 shows the data collected for each context event. Data was collected from 58 users during April 2019 to May 2020. The smartphone app listened for 144 broadcasts that occurred 16,257,795 times in 30 days for all the participants combined.

29.4.3 Experimental Setup

To conduct the empirical study, we implement majority baseline with four possible estimators (including stratified, most_frequent, prior, and uniform) and CRF using sklearn[14]. Since CRF is an improvement to Maximum-Entropy Markov model (MEMM), we exclude HMM and MEMM from this study. After pre-processing, training data consists of 142,138 instances and test data consist of 28,663 instances of context events.

Neural network classifiers were implemented using Keras [4] in Python. For neural networks, context events were converted into vector representation suitable for providing input to neural networks. Since words in the context event do not typically occur in natural language, we trained the model on our own data set using Word2Vec [5] embeddings instead of using pre-trained embeddings. For example, consider the context event ('headset', 'headset plugged') with context

Table 29.5 Support values for labels in test data

Label	0	1	2	3	4	5	6	7	8	9
Support	8860	7323	4637	4039	3493	2	1	0	0	0

category as headset and action as plugged. We obtained the embeddings for both words and took the average to find the embedding of the context event. When we tried to find the similar words related to this context event, we found the closest event (*'audio'*, *'audio_becoming_noisy'*) with a similarity measure of 25%. This event was triggered by the Android device whenever there was a change in audio source. Once we obtained the word embeddings, we split the data in the ratio of 70:10:20 as training, test, and validation set respectively. Training data was used by classifiers to learn the patterns of sequences. Validation data helped to tune the hyper-parameters. The performance of the classifier was analyzed on test data.

We predicted 40 context events in our experiments which are output labels. Exhaustive preliminary experiments were performed to find hyper-parameters that work best for sequence prediction of neural networks. The hyper-parameters tuned in this study are the number of epochs (number of iterations in which the training data is inputted to the classifier), layers, nodes in each layer, optimizer (minimizes the loss value), and batch size (number of samples propagated through the network for each epoch).

Table 29.5 shows the support of the top five and bottom five labels in the test data along with its support value representing highly imbalanced nature of our data-set. Certain context events such as connection or disconnection of internet occur more frequently than events like mounting or un-mounting of SD card.

29.5 Results and Discussion

To measure the performance, we obtain the weighted average of Precision, Recall and F_1 score for each technique across 10 runs which are reported in Table 29.6. To evaluate RQ1, which aims to provide a very basic model, a dummy classifier [14] was used. The best results are obtained from stratified estimator as indicated in Table 29.6. The stratified strategy considers the class distribution of the training data while strategies such as prior and most_frequent consider the most frequent label in the training set. Uniform strategy predicts uniformly at random.

To evaluate RQ2, we analyze the performance of classifiers using neural networks and compare with a dummy classifier. Table 29.6 shows the results obtained from LSTM and GRU classifiers across 10 runs. DNNs performs better than the baseline but in contrast to normal behavior, did not

Table 29.6 Precision, recall and F-1 score for various ML techniques

Technique	Precision	Recall	F-1
Stratified MB	12.60	4.99	6.24
LSTM (5-fold)	27.12	32.33	29.28
GRU (5-fold)	25.52	31.81	27.67
CRF	59.20	60.73	59.95

The technique CRF outperforms other techniques and hence indicated in bold.

perform better than CRF. In most of the cases, the model was not able to learn after 20–30 epochs and stopped early. The model learned the patterns but could not generalize well on our data set. We tried various combinations of hyper-parameter for LSTM and GRU and report the best one. 5-fold cross validation was performed but it improved the results by a very small margin. LSTM performed slightly better than GRU since it can learn complex dependencies from past context events. The classifiers were not able to predict 31 out of 40 labels even after using class weights which indicates that it is difficult for neural networks to learn context dependencies from our data-set. To evaluate RQ3, we implement Named Entity Recognition (NER) CRF using the Limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm. CRF outperformed other models including DNNs for c_1 value of 0.95 and c_2 value of 0.01. While CRF model tries to maximize the inference both at training and test data, it was able to learn better due to use of latent variables. These variables record the latent structure of hidden variables and observed data. Table 29.7 indicates the most frequent transitions for CRF model. We infer that it is likely that the battery level of a device increases after it is connected to power but transitions from power to other events like *date_changed* is penalized as shown in Table 29.8.

Table 29.9 represents weights of the top five and bottom five labels in our data set. We observe from the table that certain context events like (*'data_connection'*, *'wifi_connected'*) and (*'location'*, *'gps_connected'*) are remembered well by the model and are important context events. Similarly, the model does not learn certain context events such as (*'usb'*, *'device_attached'*) and (*'external_storage'*, *'media_scanner_finished'*) indicating these events may not be important.

By learning above mentioned class weights and transitions efficiently, CRF predicted the labels with F-1 score of about 60%, thereby outperforming MB and deep neural networks such as LSTM and GRU by a large margin. The main difference between NN and CRF is that neural network use a shared latent representation to learn the dependence between output variables while CRF learns these dependencies as a direct function of the output variables. So if connections are added among nodes in the output layer then sometimes hid-

Table 29.7 Most common transitions

Transitions	Weights
('power', 'power_connected') -> ('battery', 'battery_ok')	11.4185
('location', 'gps_connected') -> ('audio', 'audio_effects_open')	9.5508
('data_connection', 'wifi_disconnected') -> ('data_connection', 'lte_connected')	9.0543
('data_connection', 'lte_connected') -> ('audio', 'audio_effect_open')	9.0027
('usb', 'usb_device_attached') -> ('configuration', 'configuration_changed')	8.5492

Table 29.8 Least common transitions

Transitions	Weights
('data_connection', 'wifi_connected') -> ('headset', 'headset_plugged')	-3.2328
('date', 'date_changed') -> ('power', 'power_connected')	-3.2934
('date', 'date_changed') -> ('ringer', 'ringer_normal')	-3.3082
('power', 'power_connected') -> ('time_date', 'date_changed')	-3.8247
('time_date', 'date_changed') -> ('data_connection', 'wifi_connected')	-3.8244

Table 29.9 Most and least common events

Most frequent events		Least frequent events	
Event	Weight	Event	Weight
('data_conn', 'wifi_connected')	5.1024	('external_storage', 'media_scanner_finished')	-3.8143
('location', 'gps_activity')	4.7829	('usb', 'device_detached')	-3.5195
('ringer', 'ringer_vibrate')	4.2881	('external_storage', 'media_scanner_started')	-2.9040
('audio', 'audio_effects_opened')	3.3693	('data_connection', 'umts_connected')	-2.8997
('time', 'time-zone_changed')	3.1358	('ringer', 'ringer_normal')	-2.7860

den layer is not necessary to get good performance provided a good set of features.

29.5.1 Threats to Validity

The data obtained in this study may not be representative of all users. The distribution of data poses threats to validity as certain users have high usage in comparison to others. In addition, we tuned hyper parameters for a small data set and ran the best parameters for 100 epochs for neural networks. A larger data set will strengthen the resulting scores and may provide better accuracy.

29.6 Conclusion and Future Work

Context aware environments for smartphones, autonomous vehicles, and IoT devices pose problems for security and software testing. In the smartphone domain, user and context events create a large state space. This work examines a dataset of context events so that we may better understand real world context even patterns. A second contribution of this work is the implementation and empirical study of four ML algorithms to predict context event sequences. Result indicates that CRF outperforms other models thereby predicting context events with F-1 score of about 60%. We anticipate that future work will leverage our data, data collection framework, algorithm applications, and results in areas such as autonomous vehicles and IoT.

References

1. Android developer reference. https://developer.android.com/reference/android/content/Intent.html#constants_2. Accessed 11 August 2020
2. J. Arbon, AI for software testing, in *Pacific NW Software Quality Conference. PNSQC* (2017)
3. J. Arbon, C. Navrides, K. Toley, R. Bedino, V. Fan, M. Petersen, Abstract Intent Test (AIT) syntax (2018)
4. F. Chollet, keras (2015). <https://github.com/fchollet>
5. K.W. Church, Word2vec. *Nat. Lang. Eng.* **23**(1), 155–162 (2017)
6. D. Cutting, J. Kupiec, J. Pedersen, P. Sibun, A practical part-of-speech tagger, in *Proceedings of the Third Conference on Applied Natural Language Processing* (1992), pp. 133–140
7. F.A. Gers, E. Schmidhuber, LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Trans. Neural Netw.* **12**(6), 1333–1340 (2001)
8. T. Gogar, O. Hubacek, J. Sedivy, Deep neural networks for web page information extraction, in *IFIP International Conference on Artificial Intelligence Applications and Innovations, Artificial Intelligence Applications and Innovations, AICT*, vol. 475 (2016), pp. 154–163
9. M. Gómez, R. Rouvoy, B. Adams, L. Seinturier, Reproducing context-sensitive crashes of mobile apps using crowdsourced monitoring, in *Proceedings of the International Conference on Mobile Software Engineering and Systems* (ACM, New York, 2016), pp. 88–99
10. A. Graves, Generating sequences with recurrent neural networks (2013). Preprint. arXiv:1308.0850
11. A. Memon, I. Banerjee, B.N. Nguyen, B. Robbins, The first decade of GUI ripping: extensions, applications, and broader impacts, in *2013 20th Working Conference on Reverse Engineering* (IEEE, New York, 2013), pp. 11–20
12. Mysql. <http://dev.mysql.com/>. Accessed 11 August 2020
13. A. Onan, Classifier and feature set ensembles for web page classification. *J. Inf. Sci.* **42**(2), 150–165 (2016)
14. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
15. L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **77**(2), 257–286 (1989)

16. D. Santiago, P.J. Clarke, P. Alt, T.M. King, Abstract flow learning for web application test generation, in Proceedings of the 9th ACM SIGSOFT International Workshop on Automating TEST Case Design, Selection, and Evaluation (2018), pp. 49–55
17. Sqlite. <https://www.sqlite.org>. Accessed 11 August 2020
18. Statista: Number of mobile app downloads worldwide in 2017, 2018 and 2022 (in billions) (2017). Retrieved Nov 8, 2020 from <https://www.statista.com/statistics/271644/worldwide-free-and-paid-mobile-app-store-downloads/>
19. C. Sutton, A. McCallum, An introduction to conditional random fields. *Found. Trends Mach. Learn.* **4**(4), 267–373 (2012)

Abstract

The introduction of WebAssembly in 2017 opened a new door for performing computation in the browser at 0.9 the speed of C/C++ code (Haas et al. ACM SIGPLAN Notices 52(6), 185–200, 2017). As browsers are the most ubiquitous software, it is now possible to build universal applications that run on every machine that has a web browser installed on it. In this paper we propose a design to build web applications that take advantage of the new performance capabilities in the browser. We also implemented this design and showed that it increases the overall performance of the web applications in our experiment.

Keywords

Web applications · Web browsers · JavaScript · WebAssembly

30.1 Introduction

Web browsers are not merely for fetching and viewing HTML documents anymore. They grew in size and complexity to perform much more than that. Now, web browsers are an operating system-like software. They can execute code, manage data, and access hardware.

The idea for allowing the browser to perform computation is not new. When JavaScript was introduced to the web, the goal was to perform simple form validation [1]. As an interpreted programming language back then, it was not meant for CPU-intensive tasks. Java was brought to the web

in the form of Java Applet to take care of such tasks. However, due to security concerns, Java Applet is being avoided in the recent years. That makes JavaScript the main programming language available in the browser, or as some like to call it, the assembly of the web. Over the years more work has been assigned to JavaScript code. For example, sending http requests in the background using Ajax, encrypting/decrypting data, running games, and performing audio processing [2].

Thanks to browser native implementation of APIs that allow JavaScript to do all that. The wide use of JavaScript, even in other places other than the browser, encourages JavaScript engines vendors to work on improving its performance. That was the goal for introducing Just in Time Compilers (JIT) for JavaScript [3]. JIT compilers improved the performance of JavaScript, but not to the point where it can compete against the performance of Ahead of Time (AOT) compiled programming languages such as C/C++.

The reasons for that could be summarized in two points. First, the flexibility of JavaScript data type makes it hard to determine precisely during runtime. It is possible for an object to change its shape at any point of time during the execution. That makes it hard to optimize code that might change unexpectedly during the runtime. In addition, code optimization is an expensive process, that might be reversed, by doing deoptimization, which is even more expensive. Second, JavaScript is memory safe, so there must be some time dedicated for the garbage collector to run.

As an attempt to improve its performance, a subset of JavaScript called ASM.JS was developed. The goal was to reduce data type complexity for JavaScript engines, so the type is determined in advance. ASM.JS performs better than JavaScript, however it still underperforms when against native code [4]. Additionally, there is a file size increase of ASM.JS code compared to Vanilla JavaScript, which is a crucial part to consider when building web applications.

In the paper titled *'Bringing the Web up to Speed with WebAssembly'* which was a result of a joint research performed

J. Alamari (✉) · C. E. Chow
 Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, CO, USA
 e-mail: jalamari@uccs.edu; cchow@uccs.edu

by people from Google, Microsoft, Mozilla, and Apple, they explained the motivation and the formal semantics and of WebAssembly. They also presented some preliminary experiments with it [5].

WebAssembly is considered the first cross browser solution that was designed by major browser vendors. In addition to being universal bytecode, it is compact, safe, and fast to execute

It is worth mentioning that WebAssembly is not limited to browsers. It can run virtually everywhere. It can even run on IoT devices [6]. WASI, a WebAssembly run time, makes it possible to execute WebAssembly code on a wide range of platforms [7]. WebAssembly can help bring speed to interpreted languages such as python, and safety to fast compiled languages such as C.

30.2 Research Motivation

We believe offloading more processing to the client-side, can help reduce network traffic and improve the overall performance of web applications. Also, it helps reduce number of requests at the backend, which improves the ability to accommodate for more users. Consequently, backend operation cost for power, storage and bandwidth will be reduced.

Going back to the client side, it seems like Internet speed is never going to be fast enough. It is common to have a bad connection even if the user is in a city with a good telecommunication infrastructure, especially if they rely on cellular data. Additionally, some users may have limited data plans that they want to save. Therefore, adapting this design benefits both, clients, and service providers.

There is an extra advantage of running applications in browsers at the client side. It is that the users will be in charge of protecting their own data. Backend service providers will not have access to clients' files. Using web cryptography API, users can encrypt their data using a passphrase and push them to the backend [8]. Next time, the user can use their passphrase to decrypt data and view them from any machine.

30.3 Paper Contribution

In this paper we concentrate on how WebAssembly can help push computation to the edge of the internet. As edge computing is concerned about bringing computation closer the end users by having servers at the edge, we are here proposing bringing computation to the user's machine itself. To improve the overall performance of web applications we consider the following:

- The type of the machine the client is using. With high end client machines, we can offload more processes to the client side.

- If a user is having a low-end device, we let our backend handle the computation for them, unless they have a slow connection.
- In case of slow connection, we can transfer more of the code logic to the client side to allow more processing of their data locally, avoiding a lot of network requests.
- High speed connection with a high-end client machine, computation can happen either on the server-side or the client-side.

In our design we propose a fast method to gather all information that our algorithm needs to determine where to run the computation. This method in most cases should not incur any noticeable overhead.

30.4 Related Work

There are dozens of studies done on how we can improve the web using client-side solutions. However, based on our knowledge at the time of writing this paper, there is no work about benchmarking the client machine and decide where to run the computation.

David Herrera et al., implemented a comprehensive analysis of WebAssembly and JavaScript performance when dealing with numerical computation [4]. They ran their experiment on a wide range of devices. They showed that WebAssembly on a high-end client-devices, can achieve a better performance than native C code. According to their experiment even JavaScript can run close to the speed of native code especially on the latest JavaScript engines.

In this paper, we also evaluated WebAssembly performance and compare it to the native code. However, we were able to run standalone WebAssembly code with minimal JavaScript glue code, thanks to the latest version of Emscripten that allows us to do that. Also, when comparing WebAssembly to native code, we made sure that both codes run on the same machine with the same workload. Additionally, we had a different goal for this paper which is quickly benchmarking the client machine to see whether we can offload more computation to it or not.

Zhen et al., examined how far we can go with client-side only solution to increase the speed of mobile browsers [9]. Since the loading of resources in the browsers is the bottleneck process, they examine caching, prefetching, and speculative loading on 24 iPhones for a 1 year. They showed how client-side solution can help improve the load time of the web application, so the overall performance of the application can be improved. Caching website resources is important, but our focus on this paper is more about execution time of CPU-intensive tasks and when it can be performed at the client-side.

Table 30.1 Client machines

Device and OS	Specifications
PC running Windows 10	2.4 GHz Quad-Core intel(R) i7 and 12 GB of RAM
Mac Desktop running macOS Catalina	3.2 GHz Quad-Core Intel Core i5 and 16 GB of RAM
Galaxy S10+ running android 10.	Samsung Exynos 9820 and 8 GB of RAM.
iPhone 11 Pro Max running iOS 14.0.1	Apple A13 Bionic and 4 GB of RAM.
LGE running Android 7.1.2	Qualcomm Snapdragon 425 and 2GB of RAM.

Table 30.2 Backend machines

Device	Specifications
In Network Server running ubuntu server.	12-Core Intel(R) Xeon(R) CPU And 64 GB RAM
AWS (Amazon Web Services) EC2 XLarge Instance running ubuntu server.	8 Core vCPU and 32 GB of RAM

30.5 Methodology

To get a clear idea about how well a web application performs, we need to compare its performance the native code. In this study we are targeting applications that run on the backend versus applications that run on the browser. Therefore, it is reasonable to compare the performance of the offloaded code to the client side versus the performance of the traditional code that runs on the cloud.

The cloud code is written in C/C++ which are considered high-performance programming languages [10]. The client-side code is written in JavaScript/ASM_JS and WebAssembly.

Indubitably, the cloud is more powerful than a single client machine. However, the client machine processor is microseconds away from the file that we need to process. Transferring a file over the network takes about a couple of hundreds of milliseconds up to multiple seconds, depending on the file size and network speed.

We argue that processing files in the browser is faster, even if we use JavaScript on a low-end device running an outdated web browser, as shown in our experiment results.

A. Experimental Setup

Client Machines (Table 30.1):

Backend Machines (Table 30.2):

B. Experiment Design in Details

For this design to give a better performance, we need to gather information about clients' machines. Based on that information, we decide where to run the computation. Our heuristic approach relies heavily on what has been implemented natively in the browser.

C. General Device information

We can check for the type of the machine the client is using with *navigator.userAgent*. We also can check for the number of CPU cores and memory size that are available on the system by accessing *navigator.hardwareConcurrency* and *navigator.deviceMemory*.

This check is quick, and it incurs a negligible overhead. We considered running some code on the client's machine, to determine its capabilities, but running code may lead to slowing down the system, causing the overall performance to degrade.

D. Device free memory

The browser does a decent job in providing developers with some features about clients' machines, but we may want more information that the browser does not provide. Such information is how much memory is free to use on the device. A device could have an 8 GB of memory, but the memory could be used up by other applications.

The way we handle this problem is by using memory allocation methods available in C/C++ standards libraries. C/C++ standard libraries are fully supported by *Emscripten* compiler [11]. Since we have access to *C stdlib/malloc ()* function, we can try to allocate a large chunk of memory more than what is needed for the application to run. If the allocation succeeds, it means the user is having enough memory on the device. Based on the total size of memory, we can estimate how much memory is being used by other processes. It is important to free the memory after the allocation has succeeded. If the allocation fails, we can try to allocate a smaller chunk of memory. If the number of failed attempts exceeds three times, we could then stop transferring the code to the client side and fallback to the cloud processing. The Number of attempts can be configured by developers. This test file does not incur any overhead, it is fast and reliable.

It is worth mentioning that web browsers do not allow allocating more than 2GB of memory in advance. However, we find out that we can set `ALLOW_MEMORY_GROWTH` flag to the compiler to allow WebAssembly Module to enlarge memory gracefully as needed [12].

E. Network

Another important aspect in our design is concerned about network speed. In the case where users have high network speed, we can either fallback to traditional cloud processing or offload work to the client-side. On the other hand, in case of users with bad and unreliable connection, we should proceed with client-side processing. Transferring code logic to the client side increases load time. However, it is faster than transferring files back and forth over the network for processing.

Testing for connection speed can be done with Network Information API [13]. Unfortunately, it is not supported in all browsers by the time of writing this paper [14]. Therefore, we provide a polyfill script to handle network speed test.

Another implication here is that if the user loses connection while working on their files. Our proposed solution for this problem is to perform online check through browser APIs. Online property of navigator object can help with this. It is natively implemented in all major browsers. Its status updates whenever there is an `http` request. By simply performing a quick `http` request in the background we can get this property updated. Users should be prompted to save their work locally if they go offline. Persistent storages in the browser such as `localStorage` and `Indexed DB`, could be utilized to store user's information.

30.6 Evaluation

In this experiment we tried to evaluate every aspect of our design. We measured the time taken by our decider module. This decider module is written JavaScript and WebAssembly.

The overall time for the decider is the sum of *load time*, *WebAssembly module instantiation*, *platform info gathering*, and *network speed test*.

The evaluation is done using different network speeds and different devices with varying specifications. We used developer tools in the browser to simulate slow connection.

We also implemented multiple applications, *server-based* and *client-based*. Then, we measure the overall time taken to perform the computation for both versions.

The role of our decider module is to switch between using *server-based* or *client-based* application according to its decision. Algorithm 30.1 shows decider algorithm.

Algorithm 30.1 Decider Algorithm

```

load_Page ()
instantiate_WASM()
IF chromeBase THEN
    Access_Network_API()
ELSE
    run_Polyfill()
END IF
set_GoodConnection()
get_Memory_From_WASM()
set_GoodSpec()
IF GoodConnection THEN
    IF GoodSpec THEN
        Client-side = true
    ELSE
        Server-side = true
    END IF
ELSE
    Client-side = true;
END IF

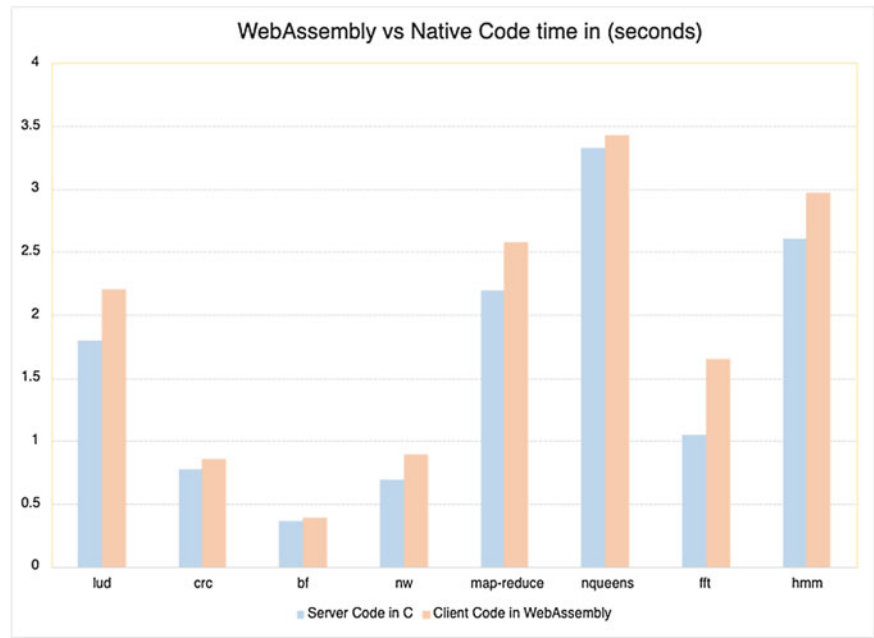
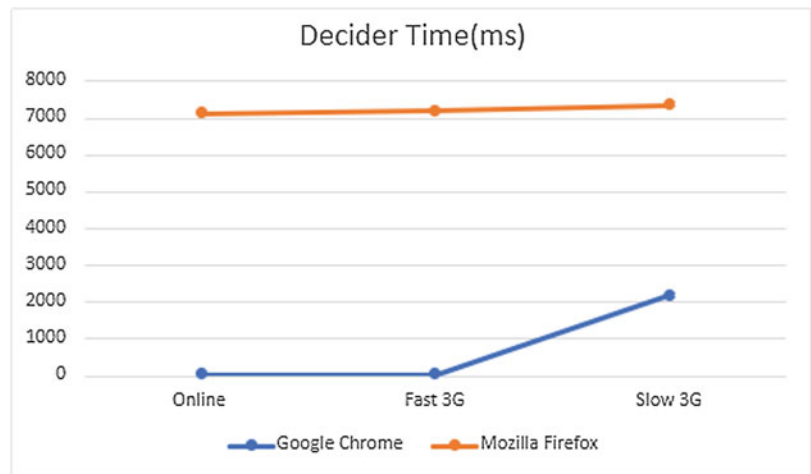
```

It is hard to test the performance of WebAssembly and compare it to native code because it runs on different machines than the one running the native code. For example, in our experiment the native code runs in a server with 12-core Xeon processor and an xlarge AWS EC2 instance.

To give WebAssembly a more level playing field we compare it against native code on the same machine. We used Ostrich Benchmark for this purpose [15]. The reason choosing Ostrich was because it supports both C and JavaScript. It does not fully support building for WebAssembly, so we had to modify it to generate WebAssembly using Emscripten compiler. We notice from Fig. 30.1 that WebAssembly performance is about 90% percent of the speed of C. The results were promising so we decided to continue the experiment and build the decider.

Figure 30.2 shows the average time taken to run the decider. As mentioned, Chrome has implemented Network Information API natively. Therefore, it takes only 16 ms to determine the speed of the network. In the worst-case scenario in our experiment, when network is set up to slow 3G, Chrome took about 2 s to get the downlink speed. The cause for this delay, is the time taken to load the script over the network since the caching was disabled.

Firefox, on the other hand, does not implement Network information API, so we had to rely on a polyfill. Our polyfill script must send chunks of data to the server and try to estimate the network speed. It is interesting that the time to do this test does not change based on the network speed in Fig. 30.1. The reason is that we are limiting the time to perform

Fig. 30.1 Ostrich benchmark performance**Fig. 30.2** Decider time

the testing. In case of good speed, we get the result in about 7 s. However, with bad connection, we do not want the code to keep waiting for the server response that might never come. Therefore, the network testing is terminated if we pass the 7 s limit. This delay time will be eliminated as soon as more browsers implement Network Information API.

Developers can be creative and show a friendly message to Firefox users telling them to wait while the application is being loaded.

After the decider is done, the decision of where to run the computation can be made either on the client-side or the server-side. We recommend running it at the client-side to avoid an extra network request.

In Fig. 30.3 we see the time taken to load and insatiate WebAssembly module.

It takes an average of 16 ms to insatiate WebAssembly module across all devices in our experiment; however, the load time is the cause for the increase of time if Fig. 30.3.

Even though WebAssembly is meant to be compact, in our experiment it was about 7 KB. It is worth motioning that this time is part of the total testing time.

In order to show real world examples of how this design could be used, we developed two applications.

First is a photo editing application. We implemented it using OpenCV library. The reason for choosing OpenCV is that it can be built for WebAssembly and JavaScript. Developers who use OpenCV.js can now use OpenCV WebAssembly version without changing their JavaScript code.

Our application has two versions. Server version written in C++ and the other version is written in JavaScript and WebAssembly.

Both codes perform the same image manipulation algorithms on the same input of a 3 MB size image.

Figure 30.4 shows the overall performance of running the code on the server side. Most of the time goes to transferring

Fig. 30.3 Time to load and instantiate WASM

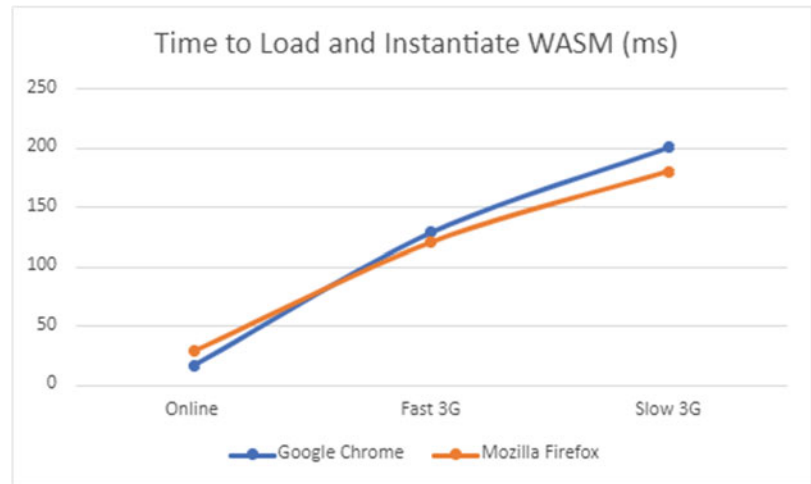
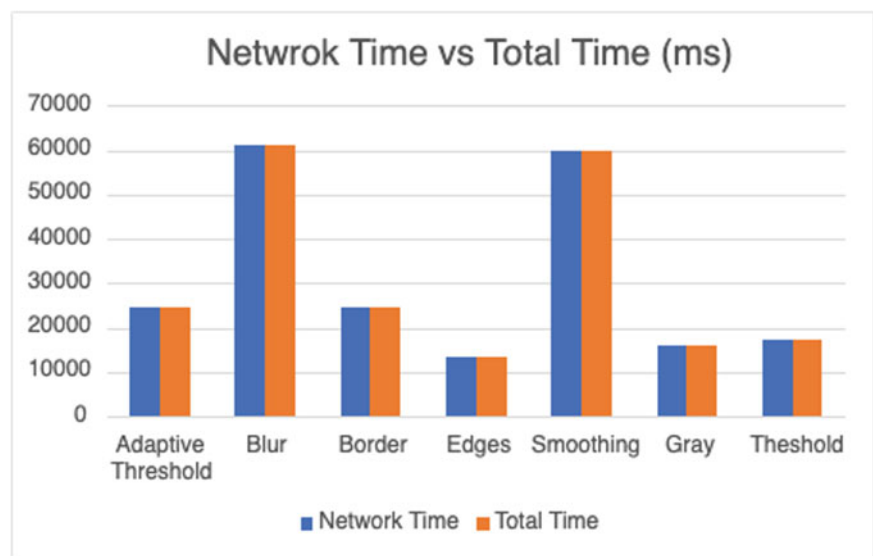


Fig. 30.4 Network time vs total time



the files over the network. The backend server is powerful, but the transfer time is the bottleneck in this graph.

If we simply switch the computation the client-side, we can improve the performance. That is clear in in Fig. 30.5. Even JavaScript code outperformed the server version, just by cutting down the network time.

Transferring source code to the client-side is not a free operation. It slows down page startup. However, source code could be cached in the browser or stored locally in an IndexedDB database. In Fig. 30.6 we see the load time of our application, with 5Mbps downlink. The load time could be shortened by more the 50% with browser caching. OpenCV code size of this experiment is 13 Megabyte for JavaScript version and 7 Megabyte for WebAssembly version. We expect load time to be longer if there are more JavaScript libraries in the project.

Another application that might benefit from this design is compression software. For this experiment we implemented two compression applications, using Zstd and Zip. Both

codes, are implemented in *C/C++* on the server, and *WebAssembly* on the client.

The reason to choose compression as an example is that it involves moving large files. Uploading large files to the server costs a huge amount of time and consumes a lot of network bandwidth. In addition, moving files to the cloud does not guarantee the safety of users' files. With this design we provide a relatively fast compression and the files never leave users' devices.

Native code that was written in *C/C++* is far more performant than our WebAssembly version as seen in Figs. 30.7 and 30.8. However, if we add transfer time of the file over the network, we end up with the numbers in Figs. 30.9 and 30.10.

In our experiment, the decider prefers running the code at the client-side with almost all configurations. The only way we got the decider to allow back-end execution was to run the code on a low-end device within the same network as our server. This will satisfy our algorithm condition of having

Fig. 30.5 WASM/JS execution time

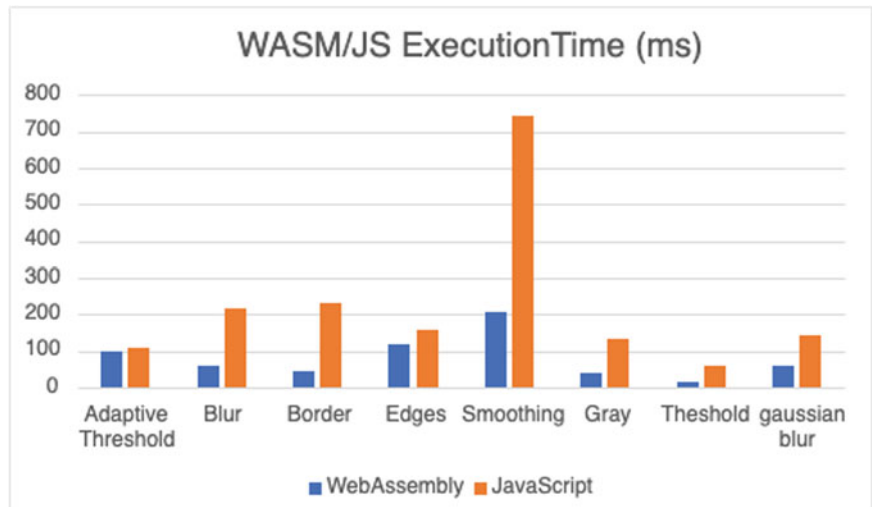


Fig. 30.6 Page load time

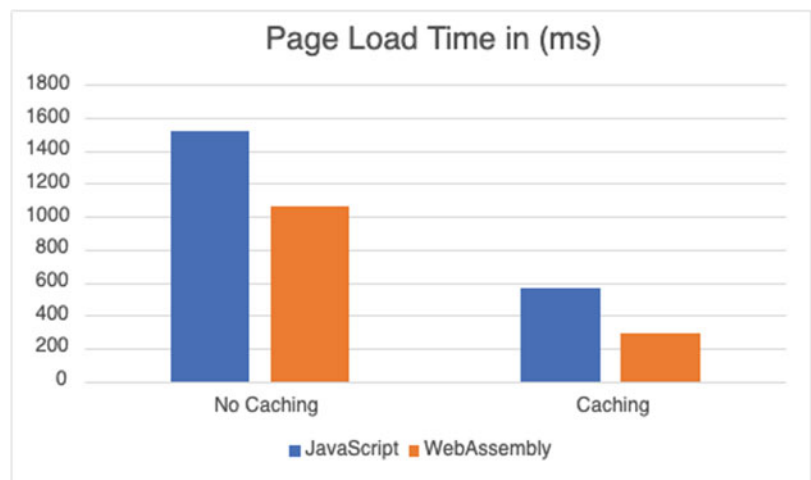
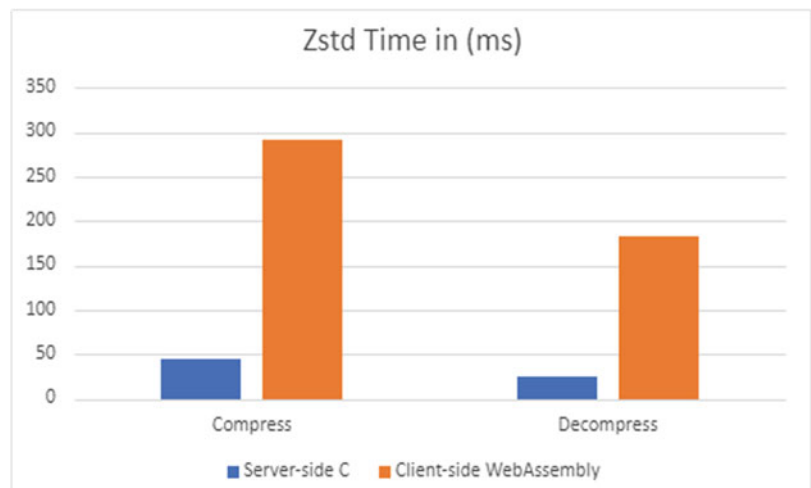


Fig. 30.7 Zstd time



a good speed and bad specifications. However, surprisingly this phone did not perform badly. OpenCV application, server version, was the same as running the application on a high-end device such as iPhone 11 Pro Max. For the client-version of OpenCV application this this phone manages to

get run code close to the speed of other devices in the experiment.

It is not surprising that the network time is the major problem when running applications that require moving large files back and forth between the client and the server.

Fig. 30.8 ZipT time

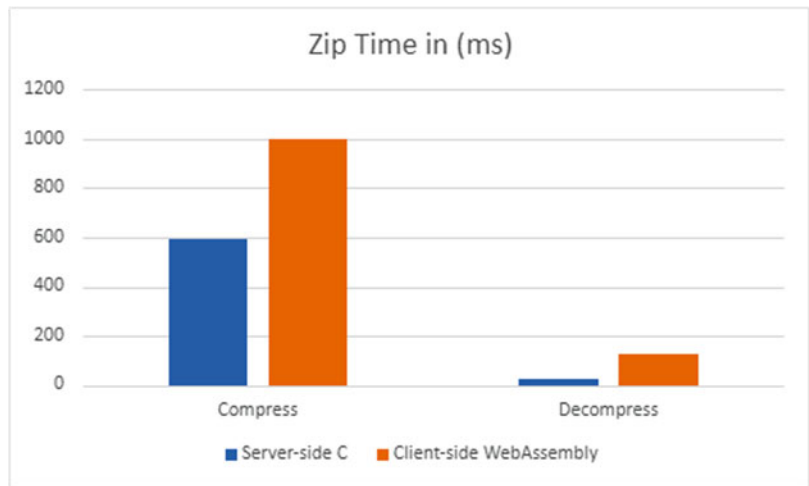


Fig. 30.9 Zstd time plus network time

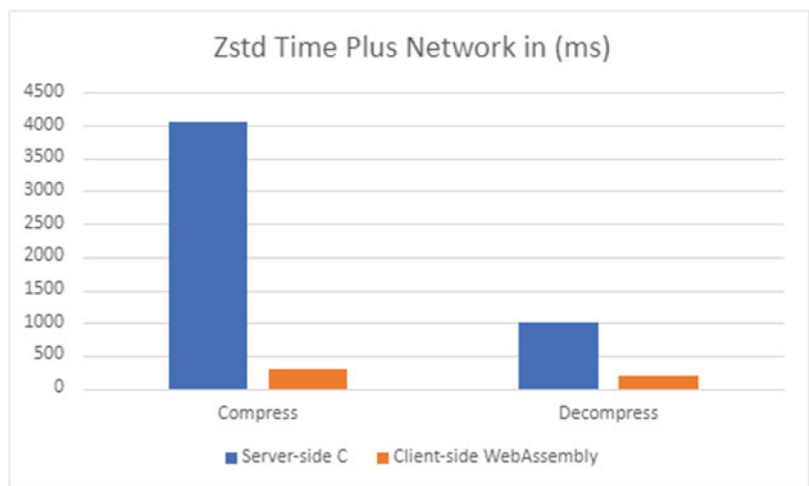


Fig. 30.10 Zip time plus network time

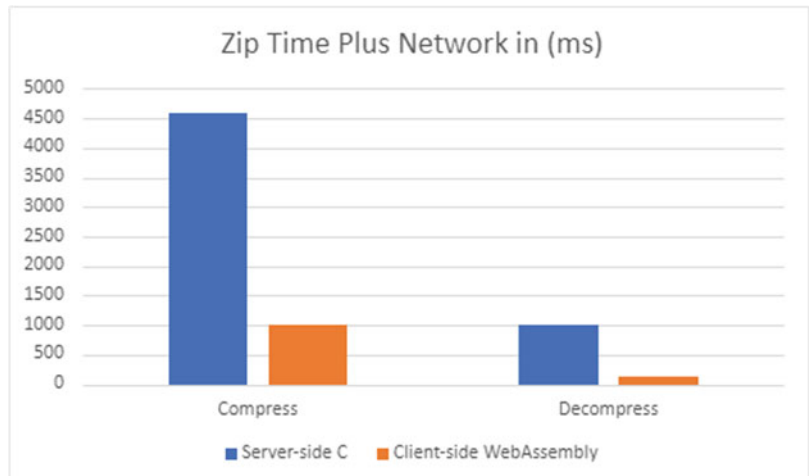


Fig. 30.11 Zstd time, LGE phone vs server

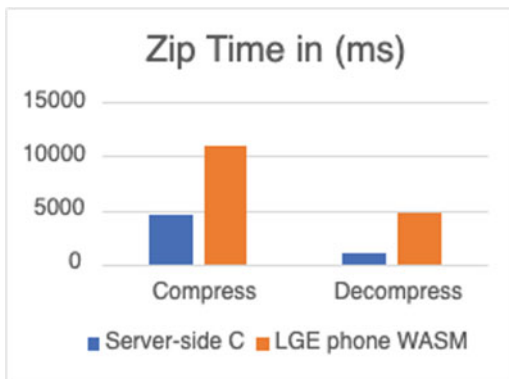
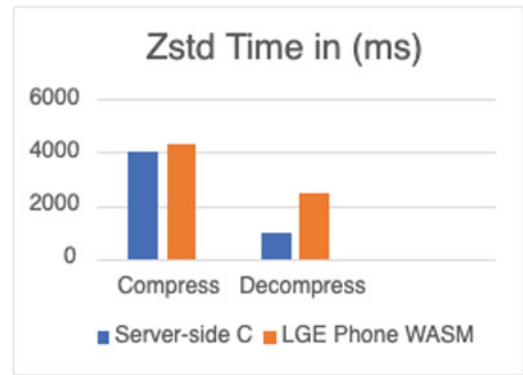


Fig. 30.12 Zip time, LGE phone vs server

This part of the experiment proves our claim about how a low-end device can outperform a powerful server just by being close to data that we need to process.

However, with this low-end device, compression application shows that keeping the computation on the back end is more efficient when it comes to performance as we can see in Figs. 30.11 and 30.12.

30.7 Future Work

We plan to further improve the decider to make a more accurate decision faster. One way to improve it could be by implementing a ML (Machine Learning) model that can recognize the device based on its user agent string without the need to perform more testing. However, this model will not be able to identify how much memory is currently available on the device nor network speed.

This decider decides where to run the whole application. It could be improved to allow developers to choose certain parts of the applications to be affected by the decision. For example, parts of the application that deals with transferring large files to the server could be performed locally based on the result of the decider and other parts of the application can continue to run on the server.

Another research idea is to study the impact of running such decider on the backend servers. For example, how much

this design can reduce backend operation cost. It would be interesting if we can find out how many cloud instances that we can be turned off by just offloading parts of the computation to the client side.

Another research direction would be working on making building C/C++ code for the web user friendly. Compiling code with Emscripten is straightforward if the project is merely using C standard libraries or some of the compiler ports that are already supported. However, building complex projects is still challenging especially if they have a lot of dependencies. We plan to build a utility that can identify code dependencies based on C source code and guide the developers through the steps required to successfully compile for the web. This tool will be more useful for JavaScript developers who have no experience with C code and *Make* building system, but they want to harness the power of WebAssembly in their projects.

30.8 Conclusion

We implemented a decider in JavaScript and WebAssembly that quickly benchmarks the client machine and determines where to run computations. This decider can be embedded in any project simply by including JavaScript file to the web page and provide a valid path to WebAssembly module. This simple yet useful decider can improve the overall performance of web applications.

As we showed in our experiments, it is feasible to build applications that are universal, performant, convenient, and secure using web technologies. We showed it is possible to push computation to the client side to benefit both parties, service providers and consumers. Clients will continue to use high quality software through the browser, and service providers will be able to provide those services at a lower cost.

Compiling large projects to WebAssembly is still a challenging task, however it is worth it on the long run. Emscripten toolchain is improving and hopefully those difficulties will be eliminated in the newer versions of the compiler.

References

1. F. Khan, V. Foley-Bourgon, S. Kathrotia, E. Lavoie, L. Hendren, Using JavaScript and WebCL for numerical computations: a comparative study of native and web technologies, in *Proceedings of the 10th ACM Symposium on Dynamic Languages, New York, NY, USA*, (2014), pp. 91–102. <https://doi.org/10.1145/2661088.2661090>
2. I. Jibaja et al., Vector parallelism in JavaScript: language and compiler support for SIMD, in *2015 International Conference on Parallel Architecture and Compilation (PACT)*, (2015), pp. 407–418. <https://doi.org/10.1109/PACT.2015.33>
3. L. Gong, M. Pradel, K. Sen, JITProf: Pinpointing JIT-unfriendly JavaScript code, in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, New York, NY, USA*, (2015), pp. 357–368. <https://doi.org/10.1145/2786805.2786831>
4. D. Herrera, H. Chen, E. Lavoie, L. Hendren, WebAssembly and JavaScript Challenge: numerical program performance using modern browser technologies and devices, p. 26
5. A. Haas et al., Bringing the web up to speed with webassembly. *ACM SIGPLAN Notices* **52**(6), 185–200 (2017)
6. M. Jacobsson, J. Wåhslén, Virtual machine execution for wearables based on WebAssembly, p. 9
7. WASI | <https://wasi.dev/>. Accessed 11 Sept 2020.
8. J. Alamari, C.E. Chow, RWA: resilient web application using client-side processing, database, and web cryptography, 2018. /paper/RWA%3A-Resilient-Web-Application-Using-Client-Side-Alamari-Chow/0e5ec005bd07a42b566cf92a16a03df270dbe48d
9. Z. Wang, F.X. Lin, L. Zhong, M. Chishtie, How far can client-only solutions go for mobile browser speed? in *Proceedings of the 21st International Conference on World Wide Web – WWW '12, Lyon, France*, (2012), p. 31. <https://doi.org/10.1145/2187836.2187842>
10. L. Gherardi, D. Brugali, D. Comotti, *A Java vs. C++ Performance Evaluation: A 3D Modeling Benchmark*, vol 7628 (Springer-Verlag, Berlin; New York, 2012)
11. A. Zakai, Emscripten: an LLVM-to-JavaScript compiler, in *Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion – SPLASH '11, Portland, Oregon, USA*, (2011), p. 301. <https://doi.org/10.1145/2048147.2048224>
12. WebAssembly. <https://webassembly.org/>. Accessed 20 Dec 2019.
13. The Network Information API. <https://www.w3.org/TR/netinfo-api/>. Accessed 12 Sept 2020.
14. Can I use... Support tables for HTML5, CSS3, etc. <https://caniuse.com/netinfo>. Accessed 12 Oct 2020.
15. *Sable/Ostrich*. Sable Research Group (2020).

Baby Shalini Ravilla, Wolfgang Bein, and Yazmin Elizabet Martinez

Abstract

The Regional Transportation Commission of Southern Nevada (RTC) manages vehicle traffic in Clark County, Nevada. The RTC division that manages the roadway and freeway street signal devices is the division of Freeway Arterial System of Transportation (FAST). They capture and store their maintenance vehicle trucks' GPS trip information into a publicly available data set. This data set brings a tremendous opportunity to analyze the traffic trends. This study uses machine learning algorithms to solve two types of problems, classification and regression. For classification, we have obtained 65% accuracy using signal types as the target and significantly low accuracy when we use the total stop time as the target variable. For regression, both Random Forest and Support Vector models performed poorly. However, the Random Forest model performed slightly better than the Support Vector model. Altogether these results help us determine the next steps to explore.

Keywords

Pattern recognition and classification · Data sciences · Traffic light coordination

31.1 Introduction

This study aims to find vehicle traffic areas in Southern Nevada and predict the peak hour based on the total stop

B. S. Ravilla (✉) · W. Bein · Y. E. Martinez
 Computer Science Department, University of Nevada, Las Vegas,
 NV, USA
 e-mail: ravilla@unlv.nevada.edu; wolfgang.bein@unlv.edu;
yazmin@unlv.nevada.edu

time metric. The data set is provided by RTC FAST and it is called “Trip status at a signal” [1]. It contains vehicle trip information captured by traffic signal devices for the year 2019. The data set contains fields such as the total stop time made by a vehicle at each traffic signal. The analysis in the sections that follow, include other important features that capture current traffic trends.

Additionally, we analyze vehicle traffic data from different signal types using classification and regression learning models. Furthermore, the results section compares the model results, given different input features. We did this to understand the vehicle traffic trend at different hours of the day along with the amount of total stop time recorded for those hours. Lastly, this study attempts to find the signal types that have more vehicle traffic accumulated.

31.2 Data Set

The data set “Trip status at a signal” used in this study was downloaded from the RTC FAST website. The data set used in the experiments section underwent a transformation via several preprocessing steps.

31.2.1 Data Preparation

The downloaded data set requires pre-processing before it can be used in the classification and regression learning models. The following is a list of pre-processing steps executed on the data set.

- Import necessary features: SignalID, Date time stamp, and Total Stop Time
- Label “Total Stop Time” as the target
- Do not import unnecessary features: latitude, longitude, corridor name, and start time

- Drop rows with missing values
- Extract “Hour of the day” and “month of the day” from the “Date time stamp” original input feature
- Load and match a second data set “Master signal ID list” [2] provided by the RTC FAST director, with the first downloaded data set “Trip status at a signal”. The second data set contains signal category types for each SignalID
- Combined data set referred to “Working File” has input features, Hour of the day, Month of the day, “signal types”, and Total stop time
- Since input features “signal types” contain string values, we performed one-hot encoding [3] to change the categorical inputs into a numerical value bit zero (off) or one (on) because machine learning models cannot accept other data type apart from numerical.
- Since input feature “Date time stamp” is cyclical, we converted the data representation to sine, cosine for the “Hour of the day” and “month of the day”. This conversion preserves the time information that hour of the day values like 23 and hour 0 are close to each other [5].

The next step in preparing the data set for the learning models is to balance the data set. We used the resampling technique [4] to balance the classes. The data set has more bias on some values, so we downsampled the data by removing the records with Total Stop Time value of zero and only consider three hundred samples from each signal type. We chose three hundred samples as a random size. Since we do not know about the size that provides an optimal solution, we tried to resample the data with different sizes and see which gives better performance metrics. In this process, the data was resampled by using downsampling, drop, and other techniques. Through a trial and error we looked for a good sample size and balance the dataset to that sample size.

The aim of this evaluation is to predict which features are giving significant results in our data analysis and what use for future studies and analysis. Using lambda, we can derive the data based on the given input size. As part of trial and error derivation of sizes, we followed another approach which is down sampling the data to a size which is expected to be considered as a marginal size for each signal type.

The last data set preparation step is to divide the data set into a train set and test set using the 80–20 split ratio. Once the data has been preprocessed, resampled, and divided, it is now ready to be fit into the classification and regression models.

31.2.2 Final Input Features

We use two pre-processed data sets. The first data set that we use has eight input features with five dependent variables from which we are going to predict the total stop time at

different hours of the day. Altogether, there are nine input features, five signal types (downtown, gateway, resort fringe, major, fire), four date/time fields (hr sine, hr cosine, month sine, month cosine), and the target output set to the Total Stop Time.

The second data set that we use has twentytwo input features which include all eighteen signal types, four date/time fields (hr sine, hr cosine, month sine, month cosine), and the target output set to the Total Stop Time.

31.3 Methodology and Experiments

To analyze the data set, we implemented machine learning models for classification and regression tasks.

31.3.1 Classification Task

For the classification task, we implemented a Support Vector Classification (SVC) model and analyze the accuracy of the train and test data set. A Support Vector Machine (SVM) is a machine learning model that is responsible for finding the decision boundary to separate different classes and maximize the margin [6]. There are two scenarios in SVM:

- SVM in linear separable case
- SVM in linear non-separable case

The latter scenario is intended for a data set that may not be linearly separable [6]. The Support Vector Classification model is implemented with the Python `sklearn.svm.SVC` library [7].

We started the classification task once the resampling of the data set was complete. We separated the data set input features into x and y clusters based on the target variable. In this scenario, we had considered two target variables as part of our experiment. We tried different input parameters and different target variables. We compared those results to find the better solution for the data set.

31.3.2 Regression Task

Since our data set is limited to a few features and some data points are biased towards a few values, this resulted in a poor classification model accuracy. Hence, we decided to try regression models.

Regression analysis helps to find the factors with the most impact and to sort out the variables that have some impact to our analysis. We refer the impacting factors as variables. There are two types of variables, dependent and independent variables.

Dependent variables are those factors with direct impact on the data evaluation whereas independent variables have an impact on dependent variables.

To perform regression analysis on the data, we consider those variables that we need to find by plotting the data and verifying the dependent variables from the entire data set. To find the relationship between dependent and independent variables we draw a regression line which is an imaginary line that runs through the middle of all data points. This line will help us to find the degree of certainty of the data prediction. Regression analysis is a basic analysis that is used to estimate the intercept, slope, and standard deviation of the errors.

There are different regression models such as linear regression, support vector, and decision tree regression. In this study, we used Random Forest regression which is an extension of the decision tree regression model [8]. Usually decision tree models are sensitive for larger training data, if the training data is changed then the result predictions may vary. Random Forest is the solution for this problem where it combines many decision trees and make into a single model. The main advantage with Random forest is that the model does not overfit and we can run as many decision trees as we want. The Random Forest regressor model was implemented using the Python sklearn library.

The approach of using SVMs to solve limitations of Regression is called a Support Vector Regression (SVR). We follow the steps below to perform SVR on our data set.

- Import necessary libraries and load the data
- Pre-process the data accordingly
- Implement model by using 3 kernels of SVR model: Linear Kernel, Polynomial kernel, and RBF kernel
- Analyze the results and see which kernel best fits the data set
- Use error metrics MSE, MAE and RMSE to evaluate the performance of the models

For the SVR kernel parameters we ran all three and reviewed the results to determine which kernel model was the best among them.

31.4 Discussion of Results

In this section we will discuss the results for the experiments of classification and regression models.

31.4.1 Classification Models

Since, we did not reach the expected performance using the Support Vector Classification (SVC) model, we decided to focus on the regression model results comparisons.

31.5 Regression Models

31.5.1 Regression Scenario 1

We compared the MSE results between Random Forest regression (RF) and Support Vector Regression (SVR) when the data was downsampled based on the total stop time 0 count and appended the data with non-zeros. By plotting the model results between actual values and predicted variables, we determined that the total stop time is mostly accumulated between 0 to 50 seconds and when compared to the SVR model, the RF model is better since more data points are accumulated within the range of 50.

31.5.2 Regression Scenario 2

We plotted the MSE actual values and predictions by considering the data set with the Total Stop Time as target variable and using inputs with signal types and hour alone as features. In this scenario, the data set that we had considered shows there is not many predicted data points that are more than 0 as total stop time. As we discussed earlier in this paper, the data is more biased towards 0 hence the actual values are not able to predict the correct values.

31.5.3 Regression Scenario 3

We also compared the MSE results from the RF and SVR models with the downsampled data with 300 sample size of each signal type and based on the type as target variable. From this scenario, we notice how the data is plotted for each signal type and which signal type has major contribution to the data set. Signal types with 2–3 have more data points that are scattered every hour whereas the actual values have a spike of data recorded at signal types 2 and 4.

31.5.4 Regression Scenario Results

We performed various experiments with data set and we tabulated the results based on the file size, downsampling the features, and selecting target variables. Even though there is minimal impact on the data size, it was significant that by changing the target variables, the MSE and accuracy values were greatly impacted.

When the number of parameters changed, the data set produced different results and they also did not have much significance on our analysis. Since the data has few limitations with the feature selection, we were unable to conduct a complete analysis and get good outcomes from the models.

The following scenarios signifies the MSE comparisons we did on different input data with different parameters considered.

31.5.5 Regression Results Scenario 1

For this scenario, the data set contained 1 million records and 22 input features which includes the types of signals along with the total stop time as a target variable. Results were tabulated by randomly selecting data of different sizes and checking their MSE values between Random Forest and SVR models.

The scenario results showed that the Random Forest model has better MSE values when compared to SVR, but the results were not significant. Hence, we tried to do different sampling of the data.

31.5.6 Regression Results Scenario 2

After considering cyclic continuous values for hour and month and adding them as one of the input features with total stop time as target variable, we get the following results. When we compare the results with earlier results (case 1), we see a significant change with the data size growing. For a larger dataset with many feature inputs, we observed a variation of MSE. This indicates that data set has more biased information and needs to be resampled.

31.5.7 Regression Results Scenario 3

By downsampling the data based on the given condition, we were able to achieve less MSE values when compared to the above scenarios. We downsampled by using the Total Stop Time count is less than or equal to 10 (i.e. we removed the records that had less stop time values). After finding a drastic change between all three cases, we tried to downsample more by removing the records that had a total stop time count under 50.

31.5.8 Regression Results Scenario 4

By changing the condition of total stop time count less than or equal to 50, below results were achieved. From the results, it is clear that by resampling the data with different input features and conditions, we can achieve better results. Even though above results has more MSE value, but it is comparatively better from all other cases.

31.5.9 Regression Results Scenario 5

In this case, we tried to classify the data set into two clusters which acts as input and target features. After classifying the data, the data set is passed to SVC model and tried to fit the data by giving type as target variable. By comparing train and test accuracy from two target variables, we can derive that for those signal types that are considered, predicted scores are almost matching to the actual value. Whereas for TotalStop-Time as target it is clearly evident that the data set has more biased information to our prediction. This proves that, the data set that has been considered is having data which has actual total stop time approximately ranging between 0 to 9000 seconds and these values are highly unpredictable to get a conclusion out of it.

To evaluate which kernel in SVM has better results, we had done the experiment on SVM by giving all three types of kernels. Below are the results for the same. Based the results, we can say linear SVM kernel works as best fit for this model among them whereas SVM-RBF kernel model performs worst. By looking at the performance of these models, the data is not performing well with either linear or non-linear kernels as MSE is too high to do any analysis on this.

31.6 Results Summary

The basic idea consists of generating a sample data set with known relevant features and irrelevant features. We had considered up to five signals types with heavy weights based on their contribution of traffic in the county. The Total stop time ranging between 0 to 9000 seconds data has been retrieved which has more data points under the 0 category. As we tried to resample the data by reducing the size of 0's total stop time records, it is still not giving enough information on how we can choose the parameters and find an optimal solution to do our analysis.

All the experiments are carried out fixing the total numbers of features. In this case we have used total number of features 8, where 7 are relevant features and 2 are irrelevant ones SIGNAL ID, Of Stops (0 1) future dive deeper here and run separate models for these majority classes. A total number of six families of data sets were generated studying two different problems (Classification and Regression).

More work to be done with classification to test more algorithms. The experiment is divided into 3 main groups. The first group explores the relationship between total stop time derived from main data and the known optimal solution. Meanwhile, the second studies the relationship of signal type with total stop time and hour, month parameters by

finding accuracy of the classification based on the target feature. Third group deals with all three types of kernels under SVM and compares them. Also consider that more than 100 data sets were generated between train and test sets in this experiment, varying the number of sample size between 5000 to 1,000,000.

31.7 Conclusion and Future Work

This work has presented various models working with different feature algorithms. The results show that based on the features and target variables model behavior is changing. We should also consider the evaluation metric results for different sample sizes.

For classification, we have obtained 65% accuracy with signal types to be the target feature whereas we got 7% accuracy for the same data with total stop time as target variable. This signifies that by considering signal type as a feature we can derive some more results and we can improve the performance of the model by considering few more input features apart from the existing ones.

For the regression type of problems, even though both the model's performance is not good, we were able to find how these two models worked better and which parameters are making some significance to the data set. Additionally, the Random Forest regression model gave better results when compared to Support Vector and thereby we can consider Random Forest and by doing some modifications on the data set we can achieve a better performance result. We had execution problems with the larger sample size of more than a million records because our server was not able to produce results because of memory allocation failures. For this reason, we failed to perform our experiments with larger size as expected.

Future work could include more feature weighting algorithms and run the experiment by varying sample size. This

work focused more on data pre-processing and generating sample data sets that gives some significant results. Another suggestion is to meet with the RTC FAST team to discuss other important features that may give a better understanding the traffic analysis.

Lastly, we were able to implement few classification and regression algorithms but there might be different approaches given the time and data set. Since the data has lots of noise and imbalanced information, even though we were able to pre-process to some extent, there is still an opportunity to change the entire data set by considering a complete analyzes strategy.

References

1. RTC FAST Signal Performance Metrics, Trips Status at Signal. <http://challenger.nvfast.org/SPM/DataDownload.aspx>. Accessed 1 Nov 2020
2. RTC FAST, Master Signal ID Labeled Category List. Accessed 1 Mar 2020
3. K.P. Murphy, *Machine Learning a Probabilistic Perspective* (MIT Press, New York, 2012)
4. S. Kota. Resampling imbalanced data and applying machine learning techniques. <https://medium.com/betterprogramming/resampling-imbalanced-data-andapplying-ml-techniques91ebce40ff4d> (2020). Accessed 1 Dec 2020.
5. C. Dossman, Top 6 errors novice machine learning engineers make (2017). <https://medium.com/ai%C2%B3-theory-practice-business/top-6-errors-novicemachinelearning-engineers-make-e82273d394db>. Accessed 1 Dec 2020.
6. L. Chen, Support vector machine — simply explained. The simplistic illustration of basic concepts in Support Vector Machine (2019). <https://towardsdatascience.com/supportvectormachine-simply-explained-fee28eba5496>. Accessed 1 Dec 2020.
7. C-Support Vector Classification Sklearn library documentation. <https://scikitlearn.org/stable/modules/generated/sklearn.svm.SVC.html>. Accessed 1 Dec 2020.
8. A. Chakure, Random Forest Regression Along with its implementation in python (2019). Retrieved from <https://medium.com/swlh/randomforestand-its-implementation-71824ced454f>. Accessed 1 Dec 2020.

Part VI

Theory and Computation

Laxmi Gewali and Samridhi Jha

Abstract

The problem of simplifying a complex shape with simpler ones is an important research area in computer science and engineering. In this paper we investigate the effect of visibility properties of polygons when their boundaries are approximated to have simpler shapes. The presented algorithm is expected to have wide applications in simplifying 1.5D terrain which are restricted class of simple polygons.

Keywords

Boundary approximation · Visibility · Terrain illumination

from some vertex of Ch_1 . The vertices of the approximated chain are usually a subset of the input chain.

The paper is organized as follows. In Sect. 2, we present an overview of important algorithms reported in computational geometry literature, dealing with polygon chain approximation. In Sect. 3, we examine the effect of polygon boundary approximation on the visibility properties of the input polygon. In particular, we show that the widely used polygon simplification algorithms do not retain visibility properties. We then present an algorithm for simplifying boundary so that the approximated polygon tends to retain visibility properties. In Sect. 4, we discuss scopes for further generalizations and investigations of the proposed problem.

32.1 Introduction

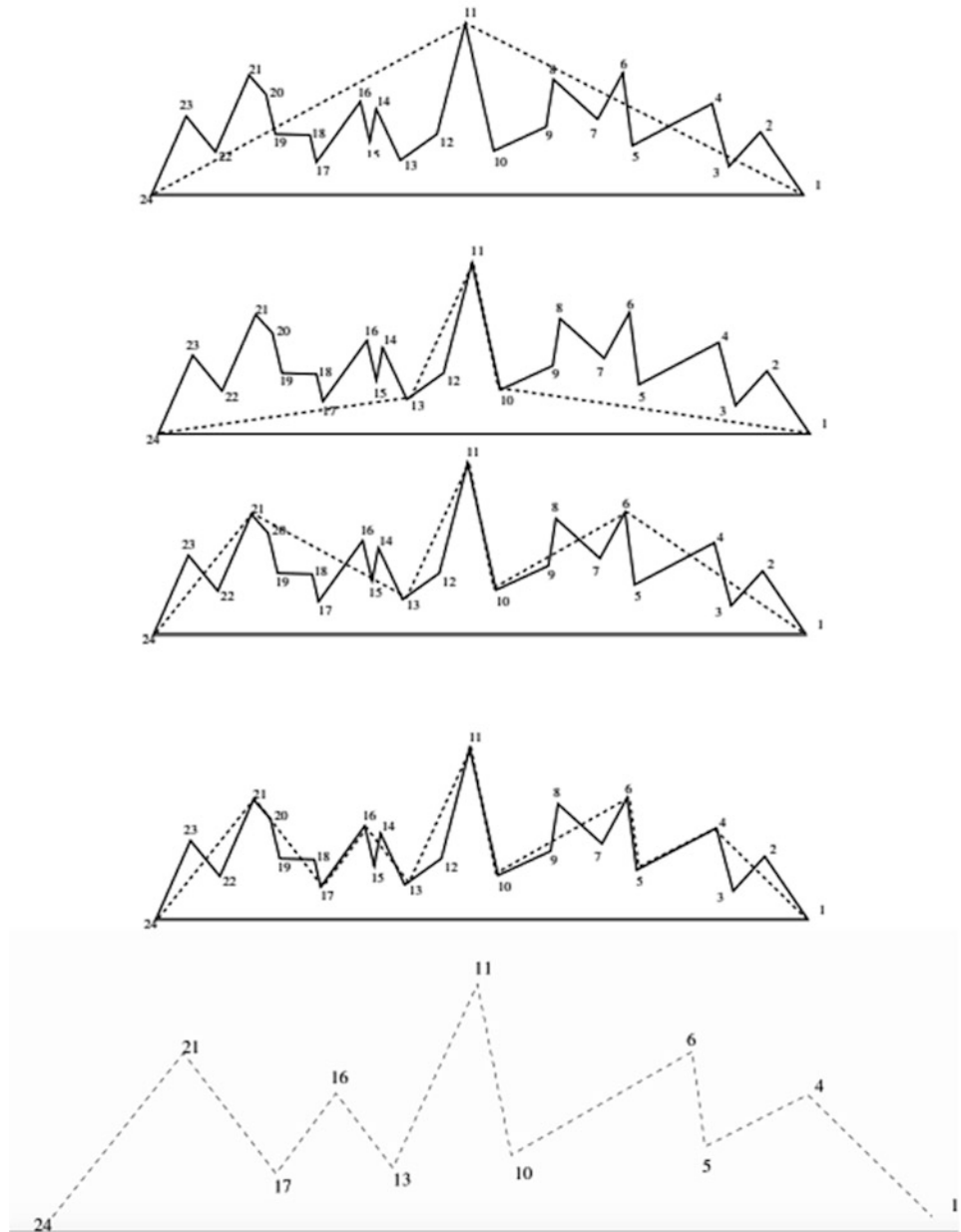
Simplifying complex polygonal chain with simpler ones has been extensively used in many application areas such as cartography, geographical information system (GIS) [1], computer graphics, medical imaging, transportation research, data transmission, and pattern recognition [2, 3]. Input polygonal chain Ch_1 is represented by listing coordinates of its vertices $p_1, p_2, p_3, \dots, p_n$ in the order they occur along the boundary. When first and last vertices are the same we have the boundary of a simple polygon. If the first and the last vertices are not same we have an open polygonal chain. Polygonal chains are approximated in term of error ϵ . The goal is to approximate the given input chain with a new one Ch_2 which has fewer number of vertices. Furthermore, each vertex of the approximated chain Ch_2 is within ϵ distance

32.2 Review of Polygonal Approximation

One of the first algorithmic attempts to approximate a complex polygonal chain $Ch_1 = p_1, p_2, p_3, \dots, p_n$ with simpler one with targeted application in cartography was investigated by Douglas and Peucker [2], which we refer to in short as D-P Algorithm. For a given error ϵ , the objective is to approximate Ch_1 by a chain Ch_2 with fewer number of vertices. Additionally, any point on the approximate chain is within distance from the original chain. D-P Algorithm takes a polygonal chain Ch_1 and tolerance error value ϵ as input and proceeds to construct approximated chain Ch_2 recursively in top down manner. We can illustrate the main idea behind D-P Algorithm by a running an example depicted in Fig. 32.1. The first part (from top) in Fig. 32.1 shows the input polygonal boundary (drawn in solid edges), where the vertices are numbered 1–24. The length of edge segment (18, 19) is taken as the value of tolerance error ϵ . D-P algorithm starts with the straight line segment $L_{1,24}$ connecting vertices 1–24 as the starting approximation for input chain Ch_1 . If all vertices of Ch_1 are within distance ϵ from the current approximation ($L_{1,24}$), then the required approximation is given by $L_{1,24}$

L. Gewali (✉) · S. Jha
 Computer Science Department, University of Nevada, Las Vegas,
 Nevada, USA
 e-mail: laxmi.gewali@unlv.edu

Fig. 32.1 Illustrating D-P algorithm



itself. In the running example, not all vertices are within distance ϵ from $L_{1,24}$. Vertex number 11 is the one furthest from $L_{1,24}$. The D-P algorithm proceeds now recursively by partitioning the input chain at the furthest vertex into two chains left chain $L_{11,24}$ and right chain $L_{1,11}$ shown by dashed edges in the first part of Fig. 32.1. The approximations of the left part and the right part are shown in the second part of Fig. 32.1. After two more rounds of recursions, all vertices are within ϵ of corresponding approximating line segments as shown in the fourth part of Fig. 32.1. The bottom most part of the figure (drawn in dashed edges) is the approximation obtained by D-P Algorithm. In this example, the D-P algorithm approximates a 24 vertices chain with a smaller

number of chain with 11 vertices. The size of approximated chain depends on the value of ϵ . For a large value of ϵ , the size of the approximated solution is small but the quality of the solution degrades. How to choose, so that the size of the approximation solution is small without compromising the quality of the solution is an important issue of D-P Algorithm.

Another algorithm for polygon boundary simplification was proposed by Imai and Iri [4], referred as **I-I** algorithm in short, works iteratively to approximate a complex polygonal chain $Ch_1 = p_1, p_2, p_3, \dots, p_n$. This algorithm first specifies a predetermined error value ϵ . The approximated chain Q is such that any original vertex is within distance ϵ from some vertex of Q . The algorithm proceeds by scanning the whole

boundary of polygonal chain. **I-I** algorithm makes use of a rectangle of width ε defines as follows

Definition 1 Given an error tolerance ε , and a polygonal chain $Ch_1 = p_1, p_2, p_3, \dots, p_j$, **ε -Rectangle** denotes the smallest rectangle of width ε that covers the maximum number of nodes in Ch_1 .

After determining the sequence of ε -Rectangles that cover all the vertices in the input polygonal chain, **I-I** algorithm proceeds to select two vertices from each rectangle to obtain the approximating line segments. The end points of approximating line segments are also the end points the corresponding sub-chains. It is remarked that the ending vertex and starting vertex corresponding to consecutive ε -Rectangles are the same. An example of the approximation process of **I-I** algorithm is depicted in Fig. 32.2. The top part of the figure shows (i) the input polygonal chain (68 vertices), and (ii) the predetermined error level ε . The middle part of the figure shows the covering of input chain with 12 ε -Rectangles. The bottom part of the figure shows the approximated chain (drawn in dotted line segments) obtained by replacing each ε -Rectangle with the corresponding approximating line segment. It is seen that a polygonal chain with 68 vertices is approximated by a simple polygonal chain with 13 vertices.

An improved version of **I-I Algorithm** is reported in [5]. This paper improves the algorithm given in [4] from $O(n^2 \log n)$ to $O(n^2)$, where n is the number of vertices in the input polygon.

32.3 Visibility Aware Approximation

In this section we present an algorithm for approximating the boundary of a simple polygon by a fewer number of vertices so that the approximated polygon tends to retain the visibility structures of the original polygon.

We start with a few technical definitions dealing with the visibility of simple polygons [6]. While studying the visibility properties of polygons, the boundary of the input polygon P is treated as an opaque wall. Two points p and q inside P are said to be **visible** if the line segment $[p, q]$ connecting p to q does not intersect with the exterior of the polygon. In Fig. 32.3, point C is visible to point D and vice versa, while point C is not visible to point E and vice versa. In term of this notion of visibility, the set of points inside the polygon visible from a given interior or boundary point R is called its **Visibility Polygon** and is denoted by $VP(R)$. In Fig. 32.3, the visibility polygon from vertex R is shown shaded. The vertices of a visibility polygon can be distinguished into two kinds: (i) *original vertices*, and (ii) *foot vertices*. While the *original vertices* are the vertices of

the input polygon, *foot vertices* are the vertices formed at the end of the chords bounding the visibility polygon. In the visibility polygon shown in Fig. 32.3, there are nine vertices, 7 of which are original vertices and the remaining 2 are foot vertices.

Let us consider the effect on visibility inside a polygon P when its boundary is approximated by a simpler polygon Q . Note that the approximation of the boundary of a polygon is essentially replacing some sub-chains with line segments.

Definition 2 Consider an instance where a small portion of the boundary $Ch_{i,j} = p_i, p_{i+1}, p_{i+2}, \dots, p_j$ of a polygon P is approximated by line segment $s_{i,j} = [p_i, p_j]$ to obtain an approximated polygon Q . An approximating line segment $s_{i,j} = [p_i, p_j]$ is said to be **visibility preserving** segment if the area of Q visible from the chain $Ch_{i,j}$ is also visible from segment $s_{i,j}$.

The above definition is elaborated in Fig. 32.4, where approximations of two chains by corresponding line segments are shown. The approximating segments are drawn in thick lines and the enclosing rectangles used for approximation are drawn in dashed lines. Let us refer the chains that are shown approximated as **left-chain** and **right-chain** with obvious meaning. Similarly, the corresponding approximating line segments are referred to as **left-E** and **right-E**, respectively. Now observe that the set of points visible from **right-chain** are also visible from the end points of **right-E**. However, not all points visible from **left-chain** are visible from **left-E**. Also observe that the shaded area is visible from the **left-chain** but not from the corresponding approximating segment **left-E**.

This motivates us to define visibility aware polygonal chain approximation problem as follows.

V-Aware Approximation Problem (VAP)

Given: (a) A simple polygon $P = p_0, p_1, p_2, \dots, p_{n1}$ (b) Error level ε .

Question Construct a simpler polygon Q with a fewer number of vertices that approximates P such that:

- (i) all vertices of P are within ε distance from some vertex in Q , and
- (ii) the visibility region of Q from any edge of Q is also visible from the corresponding chain in P .

Approximating a polygonal chain with a line segment may lead to arbitrarily large change in visibility. This is illustrated in Fig. 32.5. In Fig. 32.5, the approximating line segment $[p_4, p_{12}]$ is drawn dashed which replaces the polygonal chain $Ch_1 = p_4, p_5, \dots, p_{12}$. Before the approximation, the entire convex component $Cx_1 = [p_1, p_2, p_{14}, p_{15}, \dots, p_{16}, p_{17}, p_{18}]$

Fig. 32.2 Illustrating I-I algorithm

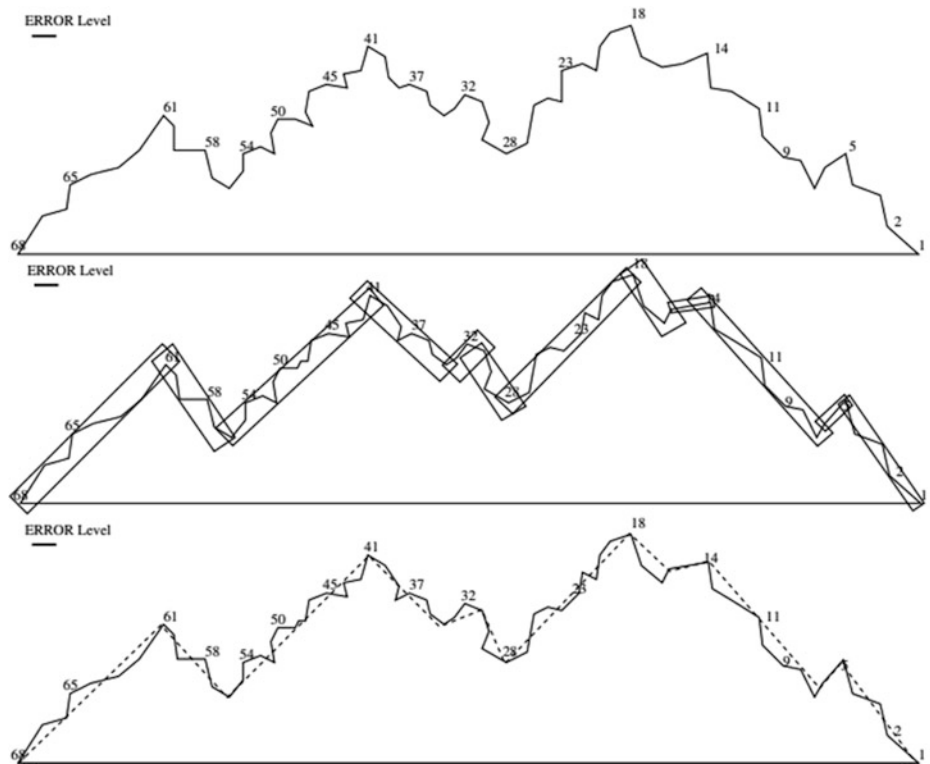


Fig. 32.3 Illustrating visibility inside a polygon

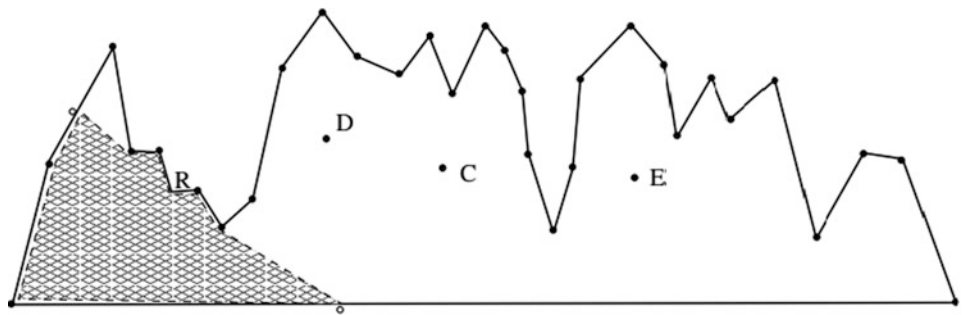
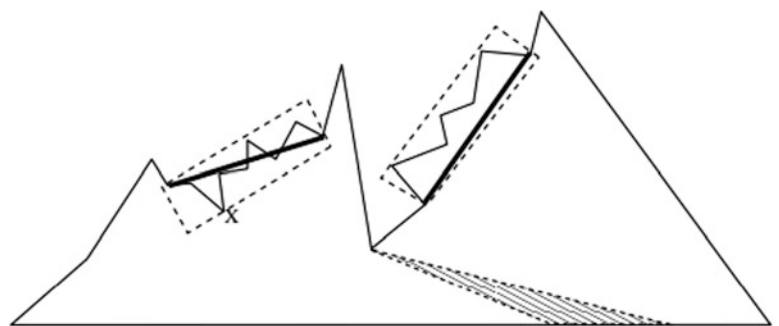


Fig. 32.4 Visibility from chain and line segment



is visible from vertex p_9 . When the chain Ch_1 is replaced by the lone segment $[p_4, p_{12}]$, none of the region inside Cx_1 is visible from the endpoints of $[p_4, p_{12}]$. This is stated in the following observation.

Observation 1 Approximating a polygonal chain by a single line segment may lead to arbitrary change in visibility.

Steiner Vertices

For mitigating the situations stated in Observation 1, we need to introduce new vertices on the edges of the input polygon which are called **Steiner Vertices**. Steiner vertices are formed by chords of the polygon constructed by connecting edges incident on reflex vertices. Figure 32.6, shows the distinction between original vertices (drawn black dots) and Steiner vertices (drawn white dots).

Fig. 32.5 Illustrating observation 1

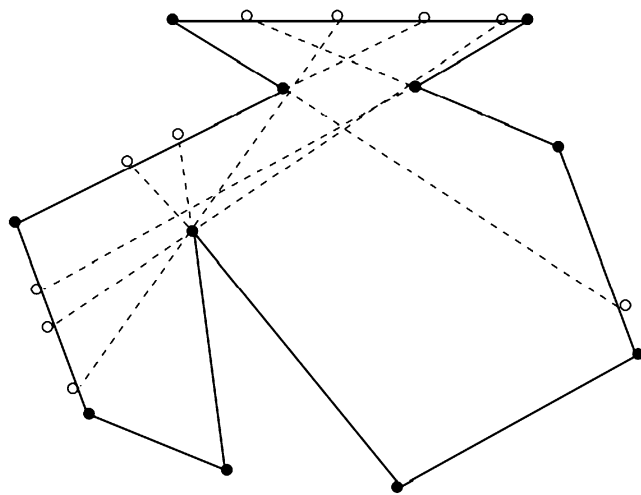
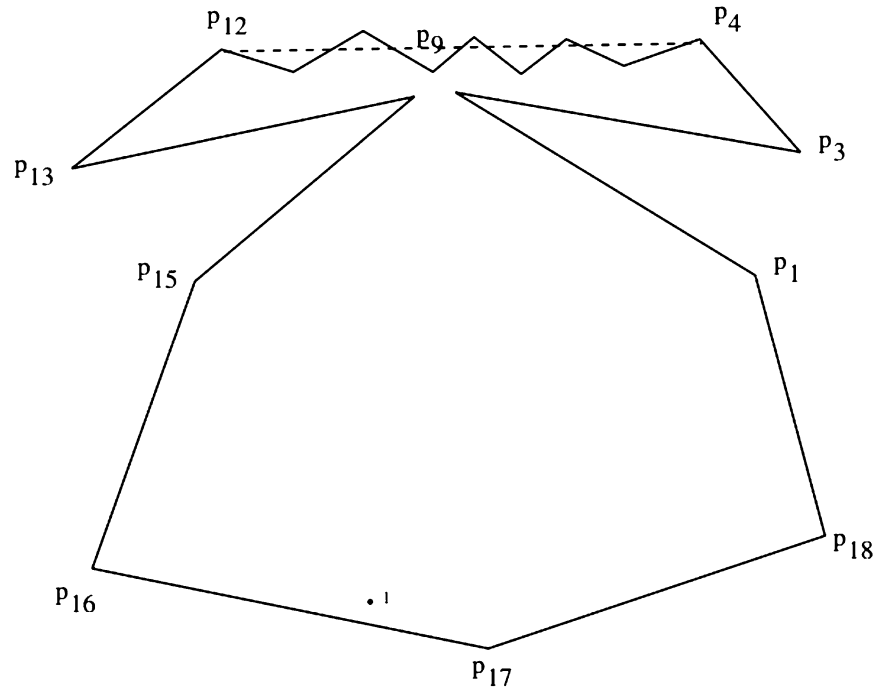


Fig. 32.6 Depicting steiner vertices

Monotone polygon and terrain modeling

A restricted class of simple polygons called **monotone polygons** have been extensively considered in computational geometry literature [6]. A simple polygon is called **monotone** with respect to a given direction d' if its boundary can be partitioned into two chains, each of which are monotone with respect to d' . Furthermore, if one of the chains of a monotone polygon is just one horizontal line segment then it is called a **monotone mountain** [6]. The polygon example shown in Fig. 32.2 is a monotone mountain. In recent years, many researchers [7] have used the term **1.5 D terrain** to indicate a monotone mountain. This term is becoming popular due to the fact that the sky line of the terrain has the form that

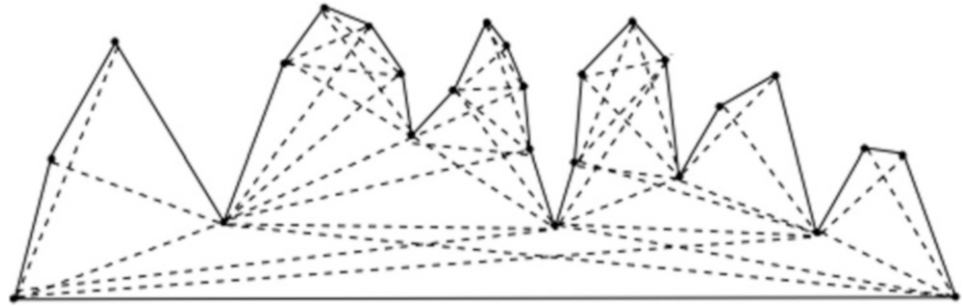
is structurally between a two dimensional terrain and one dimensional terrain.

Visibility graph of a simple polygon

A data structure extensively used for investigating visibility properties of a simple polygon is its **visibility graph** [6]. Specifically, the visibility graph $G(V,E)$ in the interior of a polygon is such that V is the set of vertices of the polygon and an edges in E consists of all pairs of vertices that are internally visible to each other. In most applications of visibility graph, **grazing visibility** is allowed which means that the boundary edges of the polygons are also part of the visibility graph. When grazing visibility is not allowed the boundary edges are not included in E . The size of the visibility graph depends on the structural shape the polygon. For convex polygon the size of the visibility graph is quadratic in the number of vertices. For polygons where large proportion of vertices are visible only with a constant number of other vertices, the size of the visibility graph is much smaller and tends to be linear in the number of vertices. Visibility graph of polygons can be computed in $|E|$ time by using the algorithm reported in [8]. An example of the visibility graph for a 1.5D terrain is shown in Fig. 32.7.

We now have technical ingredients to describe the proposed algorithm. The main part of the algorithm is the process of checking whether a given polygonal sub chain Ch_j of a simple polygon P is preserving visibility when Ch_j is approximated by a line segment L . It is noted that the approximating segment R_i is determined by finding the ϵ -**Rectangle** for Ch_j . To mitigate the problem highlighted in Observation 1,

Fig. 32.7 Visibility graph of a monotone mountain



additional vertices (**Steiner vertices**) are identified as shown in Fig. 32.6. This way, a candidate approximating-segment R_i will have needed Steiner vertices. To determine the visibility coverage of candidate-segment R_i , we compute visibility polygons from each Steiner vertices in R_i and aggregate them to determine the visibility polygon **VisiTot**. The aggregated visibility polygon **VisiTot** is compared with the combined visibility polygons of the end vertices of R_i . If **VisiTot** is contained in the combined visibility polygons of end vertices of R_i then R_i is marked as ‘visibility retaining’. It is noted that each visibility polygon is computed by navigating the visibility graph $VG(P)$ of polygon P . A formal sketch of the algorithm to determine whether R_i is “visibility retaining” or not is listed as **Algorithm 32.1**. Algorithm 32.1 makes use of function **bool VisEq(Vis1, Vis2)**, which returns **true** if **Vis1** and **Vis2** have the same set of vertices. This check is necessary to incorporate the ‘relaxed version’ of equivalency in visibility. In ‘relaxed version’ of visibility, two visibility polygons are equivalent (rather vertex equivalent) if they have the same set of non-Steiner vertices, even though they may have some uncommon areas. Algorithm 32.2 marks all visibility retaining segments of Q by repeatedly invoking **Algorithm 32.1**.

Algorithm 32.1 **bool IsCovered(P, R_i)**

```

1: Let the ordered list of vertices of  $R_i$  be  $\{q_k, p_{k+1}, \dots, p_r\}$ 
2: Compute visibility polygons  $Visi(q_k), Visi(q_{k+1}), \dots, Visi(q_r)$  from vertices  $q_k, q_{k+1}, \dots, q_r$ , respectively.
3: Set  $VisiL = Visi(q_k) \cup Visi(q_r)$ 
4: Set  $VisiTot = Visi(q_k)$ ; Set  $e = k + 1$ 
5: while  $e \leq r$  do
6:    $VisiTot = VisiTot \cup Visi(q_e)$ 
7:    $e = e + 1$ 
8: end while
9: if  $(VertexEQ(VisiTot, VisiL))$  then
10:   Return TRUE
11: else
12:   Return FALSE
13: end if

```

Algorithm 32.2 **Marking Visibility Retaining Segments**

```

1: Read (i) Polygon vertices  $P = \{p_0, p_1, \dots, p_{n-1}\}$ 
2: Read Chain vertices  $Ch_1 = \{p_i, p_{i+1}, \dots, p_j\}$ 
3: Read Error level  $\epsilon$ .
4: Compute approximating polygon  $Q$  for input polygon  $P$  using I-I algorithm
5: Let the list of approximating line segments in  $Q$  be  $R_1, R_2, R_3, \dots, R_k$ 
6: Find steiner vertices in each of  $R_i$ 's
7: for  $i = 1$  to  $k$  do
8:   if  $IsCovered(P, R_i)$  then
9:     Mark  $R_i$  Accepted
10:  end if
11: end for

```

It is remarked that the function **VertexEQ(VP_1, VP_2)**, used in Step 9 in Algorithm 32.1, returns true if visibility polygons VP_1 and VP_2 have the same set of non-Steiner vertices.

32.4 Discussion

We presented an approach for developing an algorithm for approximating the boundary of a simple input polygon P with another simple polygon Q with a fewer number of vertices so that Q satisfies two conditions: (i) each vertex of Q is within ϵ distance from some vertex in P , and (ii) most of the structural visibility of P are retained in Q . The presented algorithm is based on modifying **I-I** algorithm [4] by taking account of visibility when approximating line segment of candidate chain is computed. Preliminary investigation of the proposed algorithm shows that visibility preservation in the approximated polygon is significantly effective when the input polygon is restricted to *1.5D* terrain. Detail experimental investigation of the proposed algorithm are ongoing. These experimental results will be reported in the full version of the paper.

References

1. M. Hacı, T. Gokgoz, et al., ISPRS Int. J. Geo Inf. **8**(2), 81 (2019)
2. D. Douglas, T. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica* **10**, 112–122 (1973)
3. Z. Xie, Z. Ye, L. Wu, Research on building polygon map generalization algorithm, in *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, vol. 3, (SNPD, 2007), pp. 786–791
4. H. Imai, M. Iri, Computational-geometric methods for polygonal approximations of a curve. *Comp. Vis. Graph. Image Process.* **36**(1), 31–41 (1986)
5. W.S. Chan, F. Chin, Approximation of polygonal curves with minimum number of line segments or minimum error. *Int. J. Comput. Geom. Appl.* **6**, 59–77 (1996)
6. J. O'Rourke, *Computational Geometry in C*, 2nd edn. (Cambridge University Press, New York, 1998)
7. S. Eidenbenz, Approximation algorithms for terrain guarding. *Inf. Process. Lett.* **82**(2), 99–105 (2002)
8. S.K. Ghosh, D.M. Mount, An output sensitive algorithm for computing visibility graphs, in *28th Annual Symposium on Foundations of Computer Science*, (SFCS, 1987), pp. 11–19

A Method for Improving Memory Efficiency of the Reachability Graph Generation Process in General Petri Nets

33

Kohei Fujimori and Katsumi Wasaki

Abstract

A Petri net is a mathematical model representing the behavior of a discrete event system, and analysis of this model allows us to verify various properties of the system. When the state space is created for a net, the memory use increases when the number of states increases explosively. In this paper, we propose a method for reducing memory use during process execution by detecting and removing deletable (unused) state values among all state values in the hash map holding the state space. To judge whether a state can be deleted, it is necessary to detect routes by which the state can be reached. We developed a new method for detecting reachable paths by applying a reachability determination with a basic closed-loop matrix for the marking obtained by backward firing the transitions to the current state. A state generator was equipped with the proposed method and implemented in a Hierarchical Petri net Simulator developed at our university. We also tried to shorten the execution time by parallelizing generators equipped with the proposed method. We compared and evaluated the memory use and the execution time of the removable state deletion generator and conventional generators. The result shows suppressing runtime memory use during Petri net state space generation by detecting and deleting removable state values. Moreover, the state space generation algorithm was parallelized in an attempt to shorten execution time according to the proposed technique.

K. Fujimori
Kioxia Corporation, Tokyo, Japan
e-mail: kohei2.fujimori@kioxia.com

K. Wasaki (✉)
Faculty of Engineering, Shinshu University, Nagano, Japan
e-mail: wasaki@cs.shinshu-u.ac.jp

Keywords

Graph generation algorithm · Petri nets · Reachability graph · Memory efficiency · HiPS tool · Discrete event systems · Concurrent models · Formal verification · Basic closed-loop matrix · Paralleling generators

33.1 Introduction

A Petri net is a mathematical model representing the behavior of a discrete event system having parallelism, asynchronism, and nondeterminism of event occurrence. Analyzing the Petri net allows us to know various properties of the system [1, 2]. The Petri net design tool Hierarchical Petri net Simulator (HiPS) developed at our university has the net design, simulation capability, and various analysis functions [3–5]. A state space generator is implemented as one of the analysis functions of the HiPS. To analyze the behavior of the entire system, this generator comprehensively searches for and generates each state space as a bounded reachable graph. A state space explosion, in which the number of states explosively increases with respect to the size of the model, causes an increase in the execution time and memory use of the generator. Because resources are finite, it is necessary to reduce memory use by avoiding the generation of such an enormous state space.

During the state space generation, it is necessary to judge whether a newly generated state already exists in the previously generated state space. The state space generator can do this by determining whether the generated state value (marking vector) is registered in the hash map, the data structure that maps keys to values. Because this hash map uses a large amount of memory, sometimes the generation process cannot continue.

In this paper, we propose a method of suppressing memory use during state space generation by detecting and deleting removable state values among all the state values in the hash map. As the state space generation process progresses, the hash map accumulates state values that are not used in subsequent generation processes. Because this hash map occupies the majority of the memory used during execution of the generator, the generator memory use can be suppressed by deleting unnecessary state values from the hash map. It is assumed that the state M is a removable state value when all the next stage markings of the state M have been generated and all states of the route reaching the state M in the reachable graph have been generated. In other words, a state is removable (deletable) if it is not used in subsequent state space generation processes.

To judge whether the state M can be deleted, it is necessary to detect the routes that can reach state M . As a reachability determination method for the state, a method using the basic closed-loop matrix B_f is known [1]. We newly devised a method of detecting reachable paths to the state M by applying the reachability determination with the B_f matrix to the marking obtained by backward firing the transitions to the state M . A generator equipped with the proposed method was implemented in HiPS. We also tried to shorten the execution time by parallelizing a state generator equipped with the proposed method. We compared and evaluated the memory use and the execution time of this generator and conventional generators.

33.2 Petri Nets and Dynamic Behavioral Properties

33.2.1 Place/Transition Net

A Petri net is a model that graphically represents a discrete event system consisting of multiple processes [1, 2]. It is useful as a tool for describing and researching information processing systems characterized by parallel, asynchronous, distributed, nondeterministic, and stochastic behavior. It can simulate the behavior of a system by using tokens in the Petri net, as well as visually expressing the system structure.

The place/transition net (P/T-net), which is an ordinary Petri net, can be designed intuitively to connect components, that is, transitions, places, and arcs, with graphical notation. A formal definition of a P/T-net is given below:

A P/T-net is a 5-tuple, $PN = (P, T, F, W, M_0)$, where:

$P = \{p_1, p_2, \dots, p_m\}$ is a finite set of places,
 $T = \{t_1, t_2, \dots, t_n\}$ is a finite set of transitions,
 $F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs (flow relations),
 $W : F \rightarrow \{1, 2, 3, \dots\}$ is a weight function,

$M_0 : P \rightarrow \{0, 1, 2, 3, \dots\}$ is the initial marking,
 $P \cap T = \emptyset$, and $P \cup T \neq \emptyset$.

A Petri net structure $N = (P, T, F, W)$ without any specific initial marking is denoted by N .

A Petri net is one type of directed graph, and its initial state is called the initial marking M_0 . The basic graph N for making a Petri net is a weighted directed bipartite graph, which consists of two kinds of nodes called places and transitions. Here, arcs are either lines that connect places to transitions or connect transitions to places. Places are drawn as circles, and transitions are drawn as sticks or square nodes. The weight of a positive integer is appended to arcs, but when the weight is 1, it is often not expressed (thus, arcs without an integer are assumed to have a weight of 1). The initial marking M_0 represents the state of tokens in a Petri net in the initial state.

33.2.2 Incidence Matrix and State Equation

For a Petri net N with n transitions and m places, the incidence matrix $A = [a_{ij}]$ is an $n \times m$ matrix of integers, a_{ij} represents the number of tokens changed in place j when transition i fires once. Also, $A^- = [a_{ij}^-]$ is called a backward incidence matrix, and a_{ij}^- indicates the weight of the input arc from place p_i to transition t_j .

A marking is the enumeration of the number of tokens at all places in a Petri net. In the following explanation, a marking M_k is represented as an $m \times 1$ column vector. The j th entry of M_k denotes the number of tokens in place p_j immediately after the k th firing in a given firing sequence. The k th firing vector u_k is an $n \times 1$ column vector of $n - 1$ zeroes and one nonzero entry, and a 1 in the i th position indicates that transition i fires at the k th firing. The possible firing condition of the transition t_i is expressed by Eq. (33.1),

$$a_{ij}^- \leq M(j), j = 1, 2, \dots, m \quad (33.1)$$

and the state equation of a Petri net is defined by Eq. (33.2),

$$M_k = M_{k-1} + A^T u_k, k = 1, 2, \dots \quad (33.2)$$

33.2.3 Reachability

The marking M_n is said to be reachable from marking M_0 if there exists a firing sequence that transforms M_0 to M_n . A firing sequence is denoted by $\sigma = M_0 t_1 M_1 t_2 M_2 \dots t_n M_n$. M_n is reachable from M_0 by σ and we write $M_0 [\sigma > M_n$.

Given a Petri net (N, M_0) , from the initial marking M_0 , we can obtain as many new markings as the number of enabled transitions. From each new marking, we can again reach more markings. This process results in a tree graph of the reachable markings.

The reachability graph of a Petri net (N, M_0) is the labeled directed graph $G = (V, E)$. Its node set V is the set of all distinct labeled nodes in the reachability tree, and the arc set E is the set of arcs labeled with single transition firings such that $M_i [t_k > M_j (M_i, M_j \in V)$.

33.2.4 Necessary Condition for Reachability in an General Petri Net

Suppose that a destination marking M_d is reachable from M_0 through firing sequence $\langle u_1, u_2, \dots, u_d \rangle$. We write the state Eq. (33.2) for $i = 1, 2, \dots, d$ and sum its terms to obtain,

$$M_d = M_0 + A^T \sum_{k=1}^d u_k, \quad (33.3)$$

which can be rewritten as

$$A^T x = \Delta M,$$

where $\Delta M = M_d - M_0$ and $x = \sum_{k=1}^d u_k$. Here x is an $n \times 1$ column vector of nonnegative integers and is called the firing count vector. The i th entry of x denotes the number of times that transition i must fire to transform M_0 to M_d .

With the rank of incidence matrix A of $n \times m$ as r , A is divided according to Eq. (33.4) to make A_{12} an r -order regular matrix.

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad (33.4)$$

At this time, the basic closed-loop matrix B_f is defined as follows [1]:

$$B_f = \left[I_\mu : -A_{11}^T (A_{12}^T)^{-1} \right], \quad (33.5)$$

where I_μ is the unit matrix of $\mu = m - r$.

The condition that ΔM is orthogonal to every solution y for $Ay = 0$ is equivalent to the following condition using the B_f matrix [1]:

$$B_f \Delta M = 0. \quad (33.6)$$

If M_d is reachable from M_0 in Petri net (N, M_0) , then the corresponding firing count vector x must exist and Eq. (33.6)

must hold. Also, if Eq. (33.7) holds, M_d is not reachable from M_0 .

$$B_f \Delta M \neq 0. \quad (33.7)$$

33.3 Exhaustive State Space Generators

A state space generator is one of the analysis functions implemented in HiPS. This module generates the exhaustive behavior of the entire system as a process trace graph. By giving the pattern targeted for observation to the analysis function in the HiPS and the model checker provided as an external tool, the user can verify whether the system possesses certain properties.

33.3.1 Exhaustive State Space Generation by Single Thread

The state space in a Petri net is obtained as a bounded reachability graph. A previously proposed algorithm for bounded reachability graph generation is implemented in the HiPS state space generator [3, 6].

The generator presents all possible behaviors of the entire system as a labeled transition system (LTS), which comprises the set of all states reachable from the initial state [6]. State spaces are expressed by the LTS and output in Aldebaran automaton format [7], which is a LTS file format in the CADP Toolbox [8]. The HiPS tool also exports all possible behaviors in the Graphviz [9] DOT graph format file. The LTS format labels the transitions between states and describes the system behavior based on an event.

The single-thread search algorithm implemented in the conventional generator is Algorithm 33.1. The operation of Algorithm 33.1 is as follows: First, a firing evaluation is performed on the new marking M_k temporarily stored in *buf* using Eq. (33.1). If the evaluation indicates that firing is possible, the state transition between the next stage marking M_{k+1} , M_k and M_{k+1} , is obtained using Eq. (33.2). To generate a reachable graph, the process searches for the existence of M_{k+1} in the state space obtained so far, and if it is not included in the state space, it is registered as a new marking. Here, the information of the state space is retained by registering the generated information of all markings in the hash map (*map*). In the search for a *map* that holds the state space, if there is no state M , it is judged to be a new marking.

Algorithm 33.1 Pseudo Code of State Space Generation Algorithm (single thread)

```

1:  $buf \leftarrow M_0$ 
2:  $map \leftarrow (M_0, stateNumber \leftarrow stateNumber + 1)$ 
3: while  $buf$  is not empty do
4:    $m \leftarrow buf.pop()$ 
5:   for  $i$  in  $0..row$  do
6:     if  $t_i$  is not fireble then
7:       continue
8:     end if
9:      $m_{next} \leftarrow m + A^T u_i$ 
10:    if  $m_{next}$  is new marking then
11:       $map \leftarrow (m_{next}, stateNumber)$ 
12:       $stateNumber \leftarrow stateNumber + 1$ 
13:       $buf \leftarrow m_{next}$ 
14:    end if
15:    get transition from  $m$  to  $m_{next}$ 
16:  end for
17: end while

```

33.3.2 Exhaustive State Space Generation by Parallel Threads

To speed up generation of the state space, we implement multi-threading in the generator engine. Because the firing evaluation of the marking can be determined, generation of the next marking is completed in a finite step (Algorithm 33.1). The choice of marking from the next marking list is non-deterministic and the state space does not depend on the order of marking selection. Therefore, it is possible to parallelize the loops for selecting markings from the next marking in the generation algorithm.

This algorithm is efficiently executed in parallel on a multi-core processor. Its operation runs as follows: This algorithm introduces parallel iteration syntax *parallel_while*. Most of the state space generation processes are parallelized by executing a series of parallel firing evaluations, firing processes, and result outputs with the use of *parallel_while*.

A container holding the generated state space is exclusively referred to with the thread-safe container *concurrent_map*, so that data consistency is maintained during parallel execution. Newly generated marking(s) in the state space are held in *concurrent_buf*, and the generated state space is stored in *concurrent_map*. Each container is implemented as a thread-safe container by threading building blocks (TBB) [10]. The parallel executions operate a series of processes involving firing sequence estimation, inserting and searching new markings, and obtaining transition relations. The generating process adds a marking to the state space if it finds a new marking from the *concurrent_map*.

33.4 Proposed Method Using Removable States Detection

33.4.1 On-the-Fly Removable States Detection in State Space Generation Process

In the reachability graph generation process, for the state M where the generation of the next stage marking is completed, it is possible to judge whether the state M can be deleted by judging whether all paths that can reach M have been generated by on-the-fly. When there is a reachable state M_p and a transition t that satisfies $M_p [t > M$, then there is a feasible path back to the state M .

In the proposed method, the states M waiting for firing evaluation are taken out one by one from the queue buffer, and firing evaluation and next stage marking generation is performed. In the proposed method, removable state evaluation is performed in parallel with the state space generation processing as follows:

1. Make a reachability determination with Eq. (33.6) for the state M_p obtained by backward firing the transition t with the state M that generated the next-stage marking to return to state M .
2. Record not-yet-generated paths among the enumerated reachable routes. Because the state M can be deleted when these ungenerated paths are all generated in the subsequent state space generation, they can be deleted from the hash memory.
3. By deleting such states during the generation process by this method, hash map memory use at the time of execution can be reduced.

33.4.2 Resolution of Removable States

The state value that satisfies the conditions (Table 33.1) in the state space generation process is considered to be a state that is removable because it is not used in subsequent generation processes.

Figures 33.1 and 33.2 are examples of two situations of a reachable graph. They illustrate how to determine whether marking M_3 can be deleted during the process of graph generation which is included into its partial loop paths. The

Table 33.1 Conditions of removable state

State M is removable when it satisfies both of the following two conditions for the reachable graph:

- (a) Generation of state transitions to all subsequent markings of M are completed, and then,
 - (b) Generation of state transitions from all reachable previous markings of M to M are completed.
-

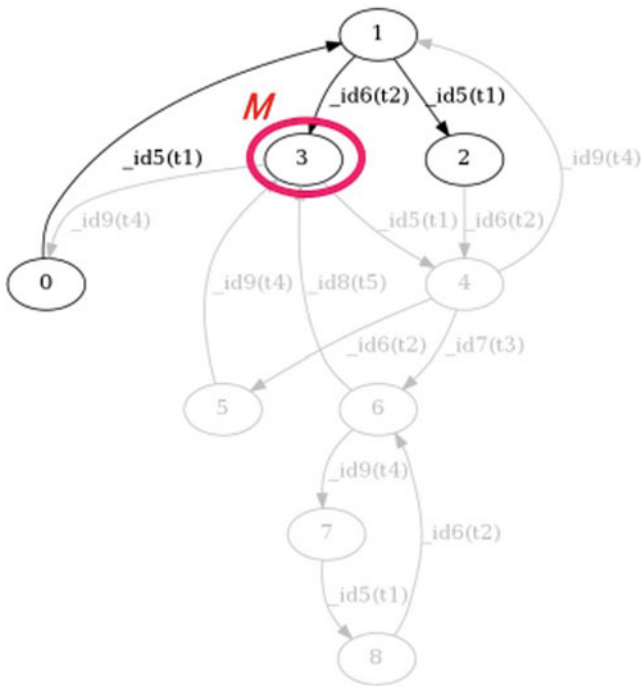


Fig. 33.1 Reachable graph up to the point of generation of M_3 . The grey arcs represent state transitions that have not yet been generated. M_3 is not removable at this stage

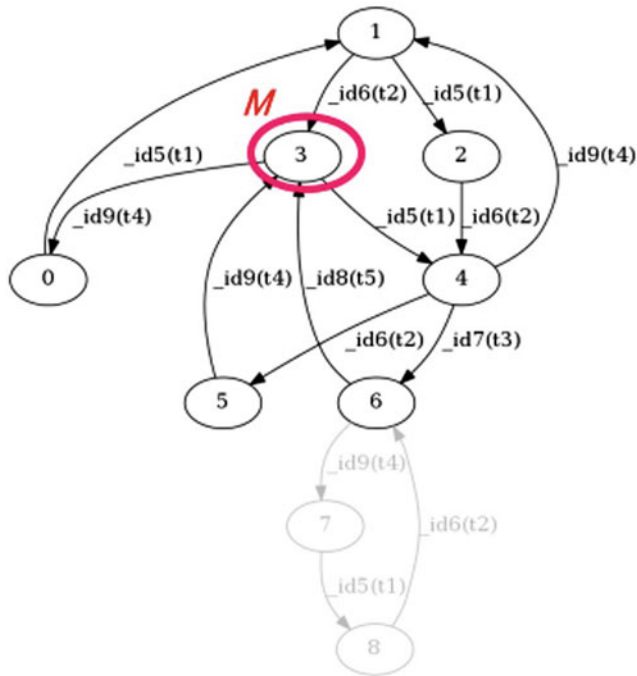


Fig. 33.2 Reachable graph immediately before generation of state M_7 . This stage at which M_3 becomes removable

creation of the reachable graph starts with the initial marking M_0 and generates the state transition related to the initial marking and the subsequent marking by prioritizing the breadth of the graph.

Figure 33.1 shows the generation process up to the point of marking M_3 . At this moment, the state transitions to the subsequent marking M_4 and M_0 from M_3 are still incomplete, so condition a) (Table 33.1) is not satisfied for M_3 . Moreover, the state transitions to the previous reachable marking M_5 and M_6 , and the subsequent marking of M_3 are also still incomplete, so condition b) (Table 33.1) is also not satisfied. In the subsequent reachable graph generation process, information about M_3 is necessary for M_3 to make the state transition to generate M_4 . Therefore, M_3 is not removable at the point in time represented by Fig. 33.1.

Next, we consider the stage at which the generation process has proceeded to just before generation of the state transition to M_7 , as shown in Fig. 33.2. At this time, all subsequent markings of M_3 is fulfilled satisfied the condition a) (Table 33.1). In addition, the state transition to the reachable previous marking M_5 and M_6 from M_3 , and the subsequent markings of M_3 have been completed, so condition b) (Table 33.1) is also satisfied.

Although the generation process continues thereafter to generate state transitions to M_7 and M_8 , the information about M_3 is no longer required in subsequent generation processes. Thus, M_3 becomes removable at the stage of Fig. 33.2.

33.4.3 Extended Algorithm for Identifying Removable Markings

To judge whether a specific marking M is removable, it is necessary to estimate the state value satisfying the conditions (Table 33.1) by on-the-fly. The generation process (Algorithm 33.1) of the conventionally registered M in the hash map (*map*) after generating all the markings at the next stage of marking M . Therefore, the finding of M through the retrieval of the map means that condition a) (Table 33.1) is satisfied.

The judgement of the condition b) (Table 33.1) depends on the list of all possible previous markings of M at a specific stage during the course of generation of the reachable graph. While the reachable graph is still being generated, the shape of the completion graph is unknown. Thus, the previous markings that could reach M are estimated with the following steps:

1. Perform reverse firing of each transition in M to obtain the previous marking candidates M' of M .
2. Apply Eq. (33.7) with $M_d = M'$ to exclude unreachable markings.

This process lists possible previous markings of M in the course of a state generation. After confirming that M satisfies both conditions (a) and (b), M is deleted from the hash map based on the judgment that it is removable.

The reachable graph generation algorithm after the addition of the judgment and deletion process is shown as Algorithm 33.2. Each state M waiting for firing evaluation is taken out from the queue buffer (buf) and firing evaluation is performed. Next, marking Eq. (33.7) is applied to the state M_p obtained by reverse firing transition t in state M , in which the generation of the next stage marking has come to completion so the route returns to state M . The unprocessed paths are recorded in the list of reachable paths. Moreover, when these uncreated paths are all generated by the subsequent state space generation, the state M satisfies the condition of both (a) and (b) (Table 33.1) and is removable from the hash map (map). In this way, by deleting unnecessary states during the generation process, it is possible to reduce memory use significantly during execution.

Algorithm 33.2 Pseudo Code for State Space Generation Algorithm with Removable State Detection (single thread)

```

1:  $buf \leftarrow M_0$ 
2:  $map \leftarrow (M_0, stateNumber \leftarrow stateNumber + 1)$ 
3: while  $buf$  is not empty do
4:    $m \leftarrow buf.pop()$ 
5:   for  $i$  in  $0..row$  do
6:     if  $t_i$  is not fireble then
7:       continue
8:     end if
9:      $m_{next} \leftarrow m + A^T u_i$ 
10:    if  $m_{next}$  is new marking then
11:       $map \leftarrow (m_{next}, stateNumber)$ 
12:       $stateNumber \leftarrow stateNumber + 1$ 
13:       $buf \leftarrow m_{next}$ 
14:    else
15:       $m_{next}.count \leftarrow m_{next}.count - 1$ 
16:      if  $m_{next}.count = 0$  then
17:         $map.erase(m_{next})$ 
18:      end if
19:    end if
20:    get transition from  $m$  to  $m_{next}$ 
21:  end for
22:  for  $i$  in  $0..row$  do
23:     $m_{prev} \leftarrow m - A^T u_i$ 
24:     $\Delta M \leftarrow m_{prev} - M_0$ 
25:    if  $B_f \Delta M = 0$  then
26:       $m.count \leftarrow m.count + 1$ 
27:    end if
28:  endfor
29:  if  $m.count = 0$  then
30:     $map.erase(m)$ 
31:  end if
32: end while

```

33.4.4 Parallelization of Detection and Deletion for Removable States

It is assumed that Algorithm 33.2 operates on a single thread. We then can apply this algorithm parallelized according to the parallel Algorithm of the conventional method. This parallelized algorithm executes a series of processes that consists of firing evaluation, next marking calculation, judgment of deletion potential, and result output. Even if insertion and deletion processing of the state value of the container comes under execution in parallel, the data integrity continues to be intact for a container that retains the state space using the thread-safe *concurrent_map*.

33.5 Performance Comparison

33.5.1 Experimental Conditions

To compare and evaluate the performance of the conventional state space generator and the generator implemented with the proposed technique, the generation process for all states was executed with a model case (TCPcondis [11]) with a partially consistent structure to make a comparisons between memory use and execution time.

For performance measurement, a physical machine and a Hyper-V virtual machine, which could assume various configurations of CPU cores and memory, were prepared. Both operating systems were Windows 10 Professional $\times 64$. Table 33.2 shows the processor performance and the memory resources of each machine used for the performance measurement.

33.5.2 TCP Connection Procedure Model (TCPcondis)

The TCPcondis [11] model is based on the Transmission Control Protocol (TCP) connection/disconnection procedure (3-way handshake) specified in RFC 793. Figure 33.3 shows

Table 33.2 Specification of machines for performance testing

Machine Name	middle-04	large-22
CPU (Intel)	Core 7-870	Xeon E5-2699 v4
Physical cores	4 core/2.93 GHz	22 core/2.20 GHz
Used threads	1-8	8-20
Memory	32 GB	256 GB

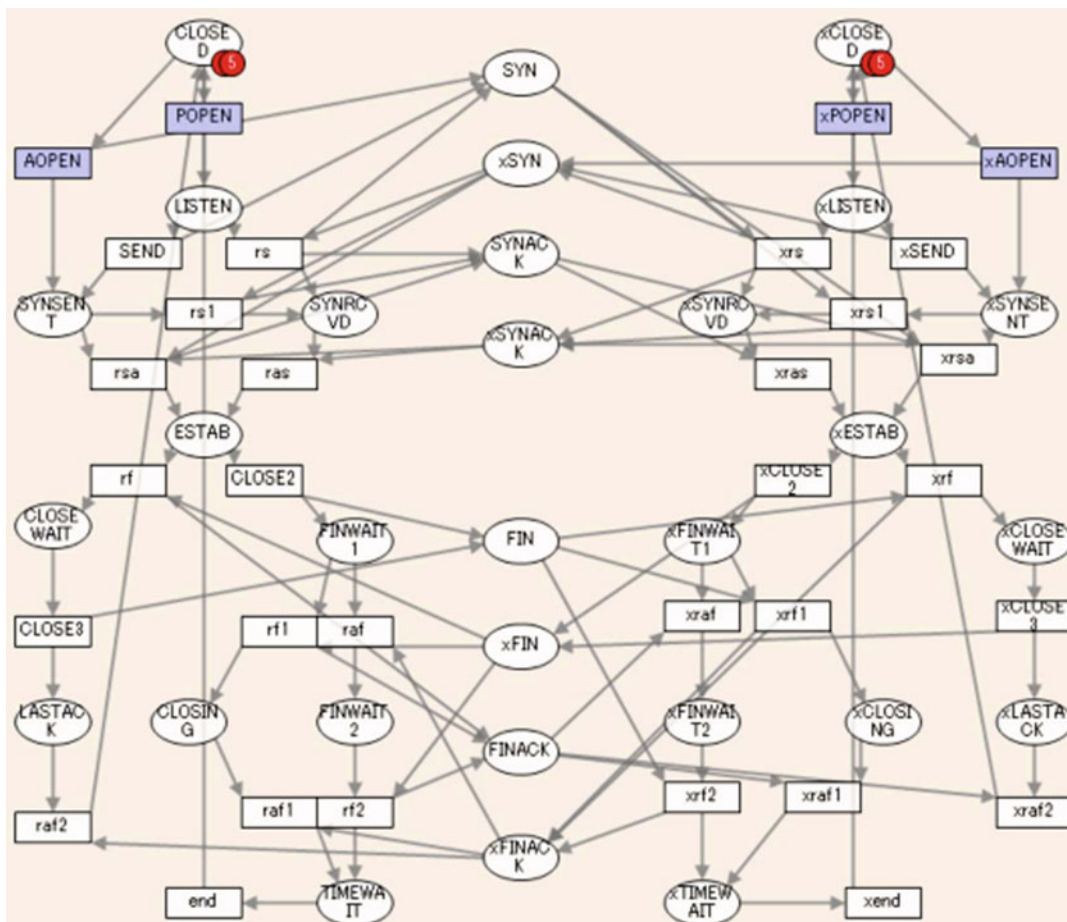


Fig. 33.3 TCPcondis 3-way handshake protocol model in TCP ($N = 5$, initial tokens on places CLOSED and xCLOSED)

Table 33.3 Size of the reachable markings and transition firings for TCPcondis

N	Number of reachable markings	Number of transition firings
1	58	112
2	1,628	6,026
3	27,645	148,534
4	329,227	2,278,628
5	2,985,834	24,899,392
6	21,803,791	210,006,762

the P/T-net model of TCPcondis. This model represents two symmetric communication nodes sharing the place with the flag information on the TCP header at the center. The number of tokens N set in the initial marking (CLOSED and xCLOSED) represents the maximum number of simultaneous connections.

Table 33.3 shows the total number of states and state transitions in the case of $N = 1, \dots, 6$. As the number of connections N increased, the size of the overall behavior of the system increased explosively.

33.5.3 Performance Comparison by Using Single/Multi-thread Implementation

The results of the memory use testing were as follows: Figs. 33.4 and 33.5 show the measurement of the instantaneous value of the runtime memory amount up to the final transition (about 25 M states and about 210 M transitions) when first marking is dependent on the simultaneous connection number $N = 5$, using single/multi-thread implementation respectively (Due to space limitations, the description of the case of $N = 6$ was omitted). Memory use in the conventional method (dotted-green line) monotonously increased because it registers all the generated state values in the hash map. In contrast, because the proposed technique sequentially deletes removable state values, memory use was continually suppressed (red line).

The amount of the memory used reached a ceiling as processing proceeded because removable state values in the proposed technique are deleted in parallel with the state generation one by one. As the generation process approached the end, the number of removable state values exceeded the

Fig. 33.4 TCPcondis $N = 5$, memory use comparison using single thread on machine middle-04, conventional and proposed algorithm

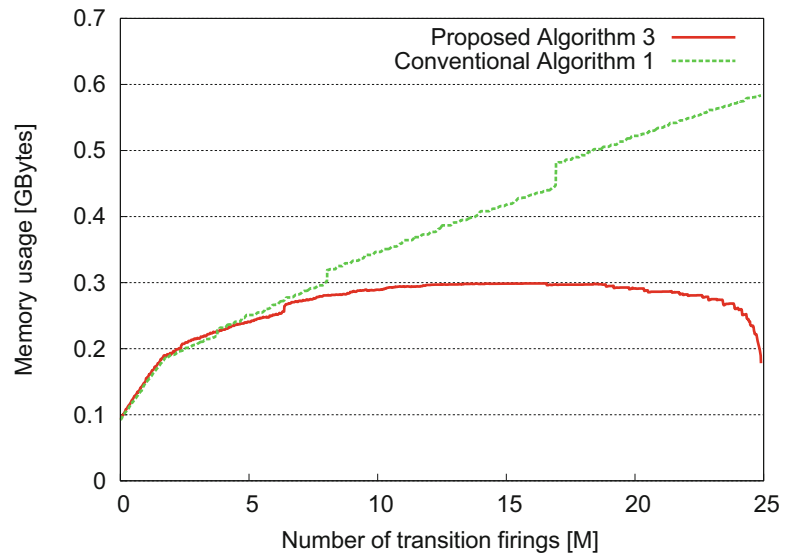
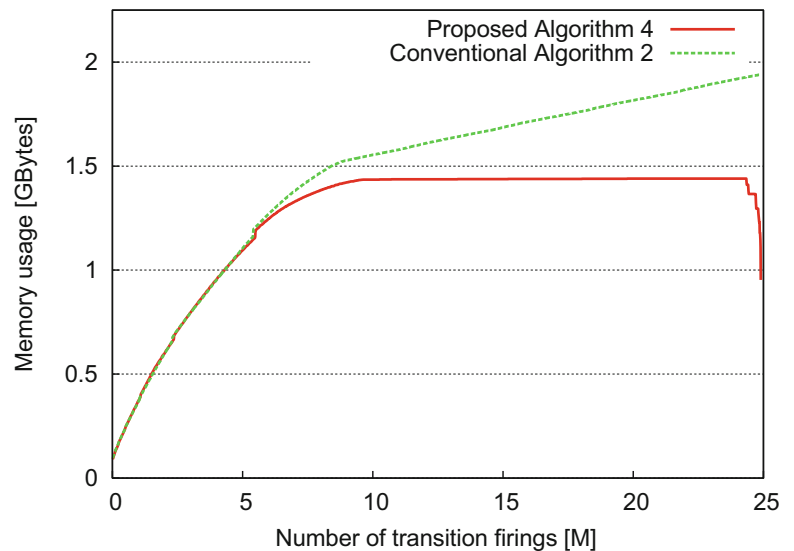


Fig. 33.5 TCPcondis $N = 5$, memory use comparison using multi-threading on machine middle-04 and large-22 (8 threads), conventional and proposed algorithm



number of states generated, and memory use was substantially reduced.

33.6 Conclusions and Future Work

The present study proposed a method for suppressing runtime memory use during Petri net state space generation by detecting and deleting removable state values. Moreover, the state space generation algorithm was parallelized in an attempt to shorten execution time according to the proposed technique. When generating a large state space with the conventional state space generator, the memory use explosively increases, making it difficult to execute the generation process.

Future tasks include the following: If it is possible to allocate considerable computational resources for pre-

processing, we can introduce a removable state value detection method using structural property analysis. For example, after obtaining the structural properties of the T-invariant as a repeated consistency in partial graphs, and during the process of detecting the removable state value.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number 19K11821.

References

1. T. Murata, Petri nets: properties, analysis and applications. *Proc. IEEE* **77**(4), 541–580 (1989)
2. C.A. Petri, W. Reisig, Petri net. *Scholarpedia* **3**(4), 6477 (2008)
3. Y. Harie, Y. Mitsui, K. Fujimori, A. Batajoo, K. Wasaki, HiPS: hierarchical petri net design, simulation, verification and model checking tool, in *Proc. of the 6th IEEE Global Conference on Consumer Electronics (GCCE)*, (2017), pp. 686–690

4. Y. Harie, K. Wasaki, A petri net design and verification platform based on the scalable and parallel architecture: HiPS, in *Proc. of the 14th International Conference on Information Technology – New Generations (ITNG2017), Advances in Intelligent Systems and Computing*, vol. 558, (Springer, 2017), pp. 265–273
5. HiPS, Hierarchical petri net simulator, Shinshu University. Available at <https://sourceforge.net/projects/hips-tools/>
6. J. Ohta, K. Wasaki, Model Designing using a petri net tool and state space generation algorithm for post-verification tool, in *Proc. of the 12th IPSJ Forum on Information Technology Conference (FIT)*, (2013), pp. 171–174
7. CADP Manual, AUT – Simple file format for labelled transition systems. Available at <http://cadp.inria.fr/man/aut.html>
8. H. Garavel, F. Lang, R. Mateescu, W. Serwe, CADP 2011: a toolbox for the construction and analysis of distributed processes. *Int. J. Software Tools Technol. Trans.* **15**(2), 89–107 (2013)
9. Graphviz, Visualization software. Available at <https://graphviz.org/>
10. TBB, Intel library – Threading building blocks. Available at <https://www.threadingbuildingblocks.org/>
11. Model checking contest (MCC’2018). Available at <https://mcc.lip6.fr/>

James Andro-Vasko and Wolfgang Bein

Abstract

Power-down mechanisms are well known and are widely used to save energy; these mechanisms are encountered on an everyday basis. We consider a device which has a set of states defined by a continuous function. The state of the device can be switched at any time. In any state the device consumes energy and a power up cost to switch the machine to full capacity. This project gives experimental results regarding power consumption to satisfy service request based on online competitive analysis. Competitive ratios, which show the effectiveness of the algorithms compare to the optimal solution.

Keywords

Online algorithms · Online competitive analysis · Power down · Renewable energy · Continuous state machine · Smart grid · Green computing · Green energy · Competitive ratio · Adaptive systems

34.1 Introduction

34.1.1 The Power Down Problem

In Information Technology energy consumption is an issue in terms of availability as well as terms of cost. According to Google [10] energy costs are often larger than hardware costs. Ways to minimize energy consumption are crucial and power usage has increasingly become a first order constraint for data centers. A growing body of work on algorithmic approaches

for energy efficiency exists, see Albers et al. for a general survey [3].

To manage power usage, power-down mechanisms are widely used: for background on algorithmic approaches to power down see [1, 2, 12, 13]. In power down problems a machine needs to be in an on-state in order to handle requests but over time it can be wasteful if a machine is on while idling. To increase efficiency devices are often designed with power saving states such as a “hibernate state”, a “suspend state” or various other hybrid states. Power down algorithms exists to control single machines or systems with multiple machines, such as in distributed machine environments.

Power-down is studied for hand-held devices, laptop computers, work stations and data centers. However, recent attention has been on power-down in the context of the smart grid [11]: Electrical energy supplied by sustainable energy sources is more unpredictable due to its dependence on the weather, for example. When renewables produce a surplus of energy, such surplus generally does not affect the operation of traditional power plants. Instead, renewables are throttled down or the surplus is simply ignored. But in the future where a majority of domestic power would be generated by renewables this is not tenable. Instead it may be the traditional power plant that will need to be throttled down.

Power-down problems are studied in the framework of online competitive analysis, see [5, 9]. In online computation, an algorithm must make decisions without knowledge of future inputs. Online algorithms can be analyzed in terms of competitiveness, a measure of performance that compares the solution obtained online with the optimal offline solution for the same problem, where the lowest possible competitiveness is best. Online competitive models the advantage that no statistical insights are needed, instead a worst case view is taken: this is appropriate as request in data centers, or short term gaps in renewable energy supply are hard to predict.

Consider a machine with a set of power states, there is an ON state an OFF state and several set of intermediate states,

J. Andro-Vasko (✉) · W. Bein
Department of Computer Science, University of Nevada, Las Vegas,
Las Vegas, NV, USA
e-mail: androvasko@unlv.nevada.edu; wolfgang.bein@unlv.edu

perhaps hibernate or a sleep state. Every state has a cost to remain in the state and a cost to power up to the ON state. A higher power state uses more energy than a lower power state but has a lower power up cost than a lower power state. We are also given a set of job requests

$$(t_1^s, t_1^e), (t_2^s, t_2^e), (t_3^s, t_3^e), \dots, (t_n^s, t_n^e)$$

where t_i^s is the when job i arrives and t_i^e is when the job completes. The jobs arrive sequentially thus, $t_i^e \leq t_{i+1}^s$ for all i . The power down problem focuses on the intervals where the machine is idle and is waiting for the next request to arrive, i.e. the time durations t_i^e to t_{i+1}^s . During this time the machine makes decisions of what state to remain in, for a discrete machine it stays in a state for some time before transitioning to lower power states, for a continuous state machine, it constantly powers down to lower power states. If a request arrives the machine needs to power up to the ON state to process the request if the machine at the time the request arrives in not in the ON state. Thus the duration of the job t_i^s to t_i^e can be neglected since power cannot be saved during this time for the power down problem, since we only focus on the time between requests. This allows us to collapse the requests into the following

$$t_1, t_2, t_3, \dots, t_n$$

where $t_i \leq t_{i+1}$. The goal is to minimize energy usage for a given set of requests by applying algorithmic techniques to schedule when the machine should transition to lower power states since the instances we choose to power down determines the energy cost. For example if we power down too soon and a request arrives early in the idle period, a favorable approach would be to remain in a higher power state. If the request arrives arbitrarily later in the idle duration, then powering down to a lower power state earlier in the idle period would yield a more favorable cost. These decisions are made in an online setting discussed in the next subsection.

34.1.2 Online Algorithms

As with any algorithm, we are given a set of input and we determine an upper bound for the algorithm. The worst case upper bound is determined by the input given by an adversary that wishes to guarantee the algorithm yields the worst case scenario which implies the maximum cost. For an online algorithm, we are given the input except the entire input is not known in advance and the algorithm must make decisions without any knowledge of future input. Once all the input has been processed, we run the exact same input against an offline algorithm. The offline algorithm knows the entire input in advance and thus can compute/yield the optimal results. Then

we compare the result to our online algorithm. Let us assume $\mathcal{A}(\sigma)$ is the cost of an online algorithm, and $\mathcal{OPT}(\sigma)$ is the cost of the offline algorithm where σ is the sequence of input and both $\mathcal{A}(\sigma)$ and $\mathcal{OPT}(\sigma)$ yields a cost to produce the input. Then we have the following

$$c \cdot \mathcal{OPT}(\sigma) \leq \mathcal{A}(\sigma)$$

where c is known as the competitive ratio, this denotes the upper bound for the algorithm \mathcal{A} with input σ . For a given problem, there may be more than 1 online algorithm, and we choose the online algorithm such that its competitive ratio is minimal. For the power down problem, the input sequence σ can be thought of a series of time instances of which the requests arrive.

The online algorithm must decide which states to utilize between requests in order to attempt to minimize its cost and subsequently the competitive ratio. The online algorithm starts in the ON state right after a request is processed and continuously transitions to lower power states throughout the idle duration, the offline algorithm however chooses a state and remains in the state until the request arrives, this yields the optimal cost. The competitive ratio is determined by the rate in which the online algorithm switches from higher power states to lower power states, since the method the online algorithm switches between states determines its idle cost

34.1.3 Prior Work

Prior work has been done on discrete power down problems ranging from few states, i.e. 2 state machine, 3 state machine, and 5 state machine, to n state machine along with taper down strategies [4,6,7,9]. Work has also been done on a continuous model where states are continuously changing as the idle duration increases [8]. In previous work on continuous state machines, we devised a set of online strategies where the rate of state transitions is based on a monotone increasing function, we used functions that grow at an exponential rate and a logarithmic rate, where the exponential strategy would transition to lower states at a slow rate and eventually towards the end of the idle period, the strategy would have the online algorithm transition at a faster rate. For logarithmic strategies the behavior was the opposite from exponential strategies.

The exponential strategy had favorable results over logarithmic for a linear system, where the power up cost and idle cost would be modeled by linear decreasing and linear increasing functions respectively. However for the system that is used in this research, the idle cost and power up cost is based on a function that increases and decreases by a polynomial function at a higher degree than linear, the logarithmic strategy had favorable results.

34.1.4 Our Contribution

In this research, we run simulations on a continuous state machine using a revised strategy, where we use a piece-wise linear function that controls the transitioning rate. In Sect. 34.2 we discuss the online strategies that are used in our experimentation, Sects. 34.3 and 34.4 we show experimental results for the strategies, and then we have our conclusions in Sect. 34.5.

34.2 Online and Offline Strategies

For our system used, we have a continuous rather than a discrete model. Thus we use a continuous function that determines the cost of utilizing any of the states along with its respective cost to power up to the ON state when a request arrives. For our system, the idle power cost curve is $a(r) = 1 - r^3$ and the power up cost curve is $d(r) = 1.5r^5$ where r is the idle time between requests, thus we can compute the idle cost at time r and we can determine the power up cost when a request arrives at time r . Since the offline algorithm knows when the request arrives, it simply chooses a starting state and remains in that state until the request arrives. The offline schedule is determined by taking the derivative of $a(r) + d(r)$

$$\text{Strategy}_{ON}(r) = \left(\frac{3r}{7.5} \right)^{\frac{1}{2}}$$

For the online algorithm, we construct a schedule that is a piece-wise linear function. This piece-wise function is constructed using 3 linear equations and we assign thresholds in which one of the piece-wise function is chosen thus modifying the rate in which the online algorithm changes between states. In order to construct this piece-wise linear function, we are first given a slope s and a duration t for which the online strategy uses the transition rate s , and the 3 linear functions are

$$f_1(r) = sr$$

$$f_2(r) = s'r + b$$

$$f_3(r) = sr + x_m - s$$

where $s' = \frac{2s(1-t)+1}{1-2t}$ and $b = t(s - s')$, we use the term middle slope to denote s' in several parts in this section as well as in the following sections. The piece-wise function is

$$\text{Strategy}_{ON} = \begin{cases} f_1(r) & \text{if } r \leq x_1 \\ f_2(r) & \text{if } x_1 < r \leq x_2 \\ f_3(r) & \text{if } x_2 < r < x_m \\ 1 & \text{if } r \geq x_m \end{cases}$$

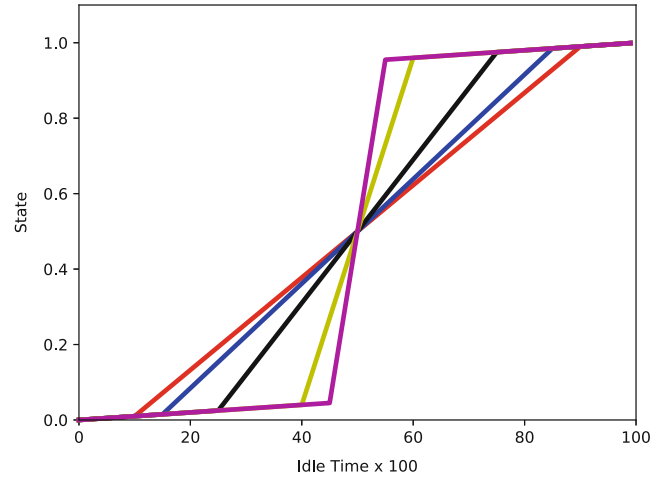


Fig. 34.1 Online strategies for slope 0.1 with durations 0.10 (red color), 0.15 (blue color), 0.25 (black color), 0.40 (olive green color), and 0.45 (violet color)

The length of the duration from 0 to x_1 is t , the duration from x_2 to x_m is also t , and the duration length of x_1 to x_2 is $x_m - 2t$. The idle duration ranges from time 0 to $x_m = 1$, at this time or later the online algorithm transitions to the OFF state and the offline algorithm would simply choose the OFF state. It is proven from prior work that the online and offline algorithms use the same threshold value that determines when the machine must choose to turn itself off in order to minimize the competitive ratio. The value of $a(r)$ at time $x_m = 1$ yields the value 0 which denotes the machine would be in the OFF state since idle time cost would be 0.

In the next sections of this paper, we construct several schedules using a set of values for s and t . Then we fix one of these values and iterate through the other to determine patterns for the competitive ratio

34.3 Analysis of Changing Duration with a Fixed Slope

In this section, we choose two extreme slopes, an arbitrarily small and large slope s , and we iterate through a set of durations t . We use a set of schedules for the online algorithm to determine the state in which it using at time r where each schedule uses a different s and t value for its given input. In this section we choose two extreme slope values, $s = 0.1$ and $s = 0.75$ and we assign a set of durations t for which the online strategy uses for the length for which the slope is 0.1 and 0.75.

Figure 34.1 shows 5 strategies, each strategy uses a slope of 0.10 for the rate of transitioning from high power states to lower power states from time to x_1 and time x_2 to x_m , thus the length of which this transition rate s is used by the

Table 34.1 Results of strategies when $s = 0.1$

Duration	Middle slope	Comp ratio
0.10	1.225	2.591
0.15	1.386	2.646
0.25	1.900	2.787
0.40	4.600	3.108
0.45	9.100	3.259

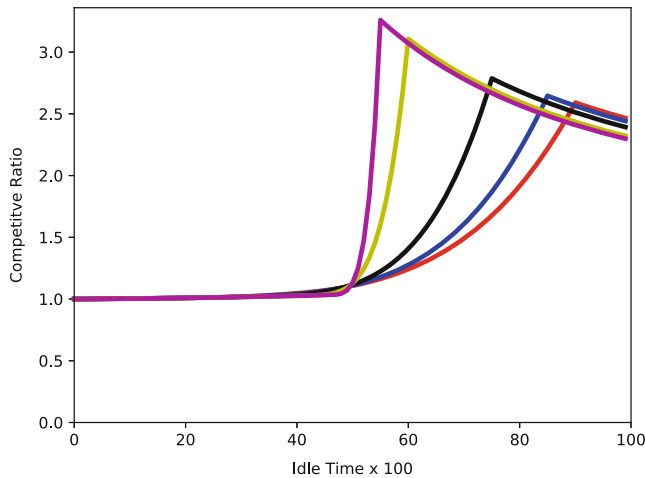


Fig. 34.2 Competitive ratio for slope 0.1 with durations 0.10 (red color), 0.15 (blue color), 0.25 (black color), 0.40 (olive green color), and 0.45 (violet color)

online algorithm varies for each strategy. We see that when the strategy that has a larger t value, for example when $t = 0.45$ then the middle slope is larger than for example when $t = 0.1$, but the duration of the middle function is shortened. This can be seen in Table 34.1. When the strategy uses an arbitrarily faster transition rate, the online cost increases at a larger rate than the other functions with a smaller t value.

In Fig. 34.2, the possible competitive ratios throughout the idle time for all 5 strategies is shown. For each strategy, the competitive ratio is close to 1 which implies that the strategies in the earlier stages of the idle duration is minimal. Then the competitive ratio of each strategy increases after the midway of the idle duration when the transition rate increases. When we see the curves in Fig. 34.2, the overall competitive ratio for each strategy is the max value of the entire curve.

Thus, the peak value for each strategy is the max competitive ratio for the strategy, since we wish to bound the competitive ratio with the maximum value which yields the competitive ratio for each strategy. These results along with the middle slope values are seen in Table 34.1. Thus, the worst strategy of the 5 strategies is when t contains a larger value, since its peak competitive ratio value is the largest of the 5 strategies.

The strategy that uses the longest t value in which the transition rate is s , incurs the largest overall cost due to the

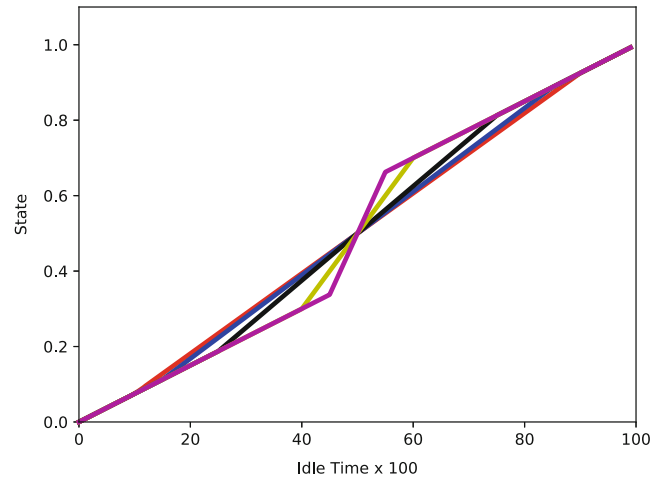


Fig. 34.3 Online strategies for slope 0.75 with durations 0.10 (red color), 0.15 (blue color), 0.25 (black color), 0.40 (olive green color), and 0.45 (violet color)

fact that the middle slope transition rate is the largest of all strategies and the power up cost increases at a higher rate which incurs a larger overall cost. Thus, the conclusion in this experiment shows the the competitive ratio is minimal when the overall transition rate throughout the idle period is a lower slope.

Now, let us investigate a set of online strategies for a larger fixed slope and each strategy once again chooses a different t value in which it utilizes the smaller slope transition rate. We utilize the same set of t values for each strategy as in the previous experiment. We see in Fig. 34.3, that each strategy has a similar rate in which the states would transition. Once again, the strategy with the largest t value has the largest middle slope which can be seen in Table 34.2.

The competitive ratios for every instance in the idle duration can be seen in Fig. 34.4, and the strategy with the largest t value has the largest intermediate peak value, however, all the strategies converge to approximately the same competitive ratio. As before however, the competitive ratios for all 5 strategies are favorable in the early stages of the idle time and then all 5 strategies converge to their maximal competitive ratio value.

In this scenario, each strategy transitions relatively at the same rate, and the s' is has a slightly larger value than the initial and ending transition rate s . Due to this, the competitive ratio for each strategy is approximately the same throughout the idle time. However the strategy that uses the largest t value has a larger peak competitive ratio value at roughly midway through the idle duration due to its larger middle transition rate s' , thus the conclusion once again is the strategy with arbitrarily smaller t values yields the overall favorable competitive ratio.

In conclusion, for this analysis, the strategy with favorable results whether the s value is arbitrarily larger or smaller, the

Table 34.2 Results of strategies when $s = 0.75$

Duration	Middle slope	Comp ratio
0.10	1.063	2.449
0.15	1.107	2.443
0.25	1.250	2.434
0.40	2.000	2.423
0.45	3.250	2.420

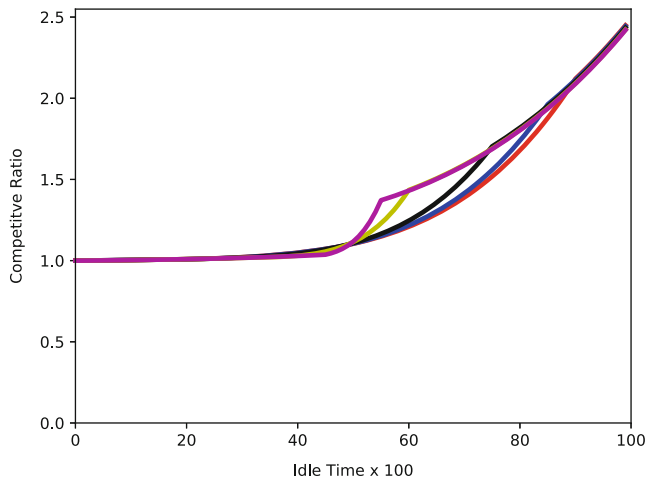


Fig. 34.4 Competitive ratio for slope 0.75 with durations 0.10 (red color), 0.15 (blue color), 0.25 (black color), 0.40 (olive green color), and 0.45 (violet color)

strategy with the smaller t value results in a more favorable competitive ratio. This was because when t is larger, the transition rate is low for a larger amount of the idle duration than when t is smaller, and thus the strategy needs to increase the transition rate to an arbitrarily large rate in order to reach the OFF state at time x_m .

34.4 Analysis of Changing Slope with a Fixed Duration

In this analysis, we have a set of 5 online strategies where the t values are fixed and we iterate through a set of slopes s . As in the previous section, we choose two extreme t values and a set of s values. The strategies are shown in Fig. 34.5.

Once again, each strategy transitions from higher to lower power states at the rate of slope s for duration t , then each switch to a larger middle slope and then they switch back to s until time x_m in which the strategy chooses the OFF state. The middle slope values and competitive ratios can be seen in Table 34.3.

From Fig. 34.6, we see the competitive ratios for the 5 strategies throughout the idle duration and the max value for each curve denotes the competitive ratio for the strategy. Once again, we see that each strategy has a favorable com-

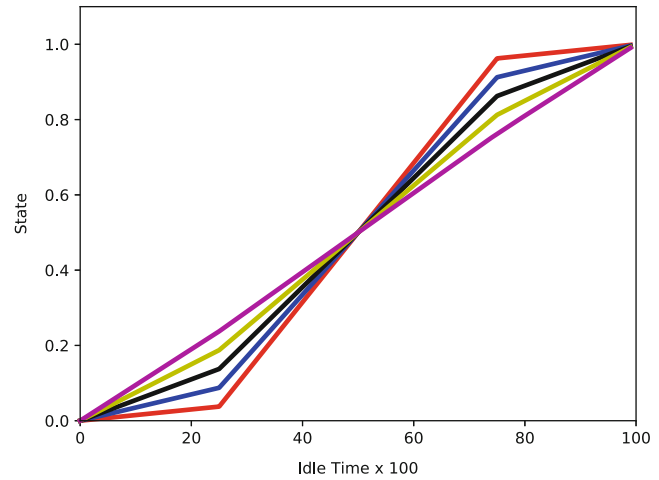


Fig. 34.5 Online strategies for duration $t = 0.25$ with slopes 0.15 (red color), 0.25 (blue color), 0.55 (black color), 0.75 (olive green color), and 0.95 (violet color)

Table 34.3 Results for the strategies with duration $t = 0.25$

Slope	Middle slope	Comp ratio
0.15	1.850	2.673
0.25	1.750	2.463
0.55	1.450	2.424
0.75	1.250	2.434
0.95	1.050	2.440

petitive ratio for first half of the idle duration, and then as they begin transitioning to lower power states at a faster rate the competitive ratio increases, and then they all switch back to the original transition rate s and they either decrease the competitive ratio or the competitive ratio increases at a slower rate.

We can see that the competitive ratio for the online strategy with the smallest s increases at a larger rate and has the largest competitive ratio. This is due to its middle slope value at which it transitions to lower power states is larger than all the other strategies, and this causes its cost to increase at a larger rate than the other strategies. Also, we can see that the strategy with the larger s transition rate which has the smallest middle slope yields the minimal competitive ratio among the strategies until the idle time x_m is reached, the competitive ratio of this strategy yields the largest. However, this would only be an issue if every request arrives at time x_m or after, thus as long as the machine is not powered to the OFF state, the strategy with the largest s value overall is the better option among the strategies.

For the last experiment, we iterate through the same s values as in the above example however now we choose $t = 0.4$. Each strategy uses the s transition rate for a longer period of time and the duration in which the middle slope is

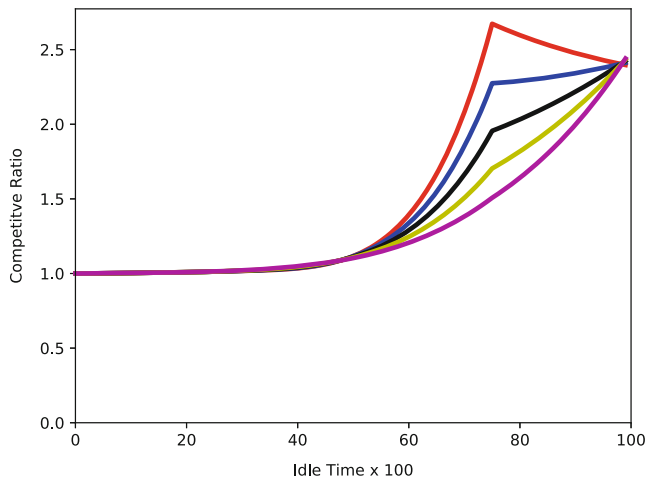


Fig. 34.6 Competitive ratio for duration $t = 0.25$ with slopes 0.15 (red color), 0.25 (blue color), 0.55 (black color), 0.75 (olive green color), and 0.95 (violet color)

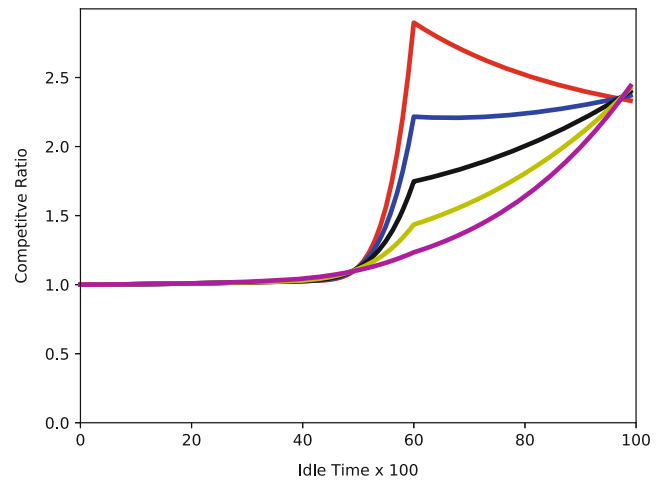


Fig. 34.8 Competitive ratio for duration $t = 0.25$ with slopes 0.15 (red color), 0.25 (blue color), 0.55 (black color), 0.75 (olive green color), and 0.95 (violet color)

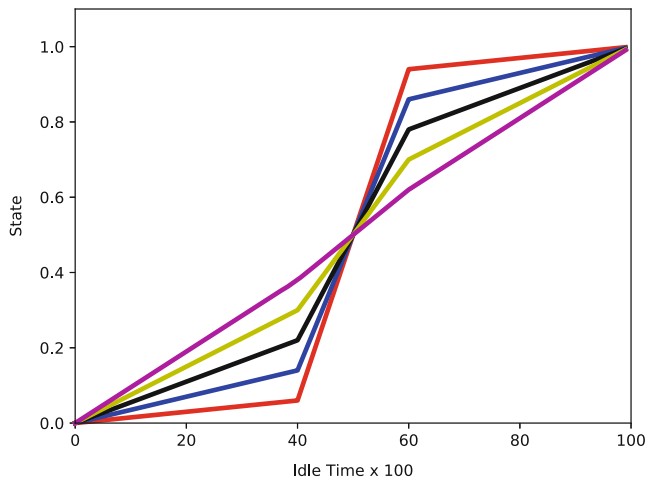


Fig. 34.7 Online strategies for duration $t = 0.4$ with slopes 0.15 (red color), 0.25 (blue color), 0.55 (black color), 0.75 (olive green color), and 0.95 (violet color)

Table 34.4 Results for the strategies with duration $t = 0.4$

Slope	Middle slope	Comp ratio
0.15	4.400	2.897
0.35	3.600	2.370
0.55	2.800	2.400
0.75	2.000	2.423
0.95	1.200	2.439

used for transitioning is decreased. We can see the 5 strategies in Fig. 34.7.

As in the previous experiment, the middle slopes decrease for strategies that use a larger s value, and as in the last example the competitive ratio is minimal when the s value used by the strategy is larger. Table 34.4.

We can see from Fig. 34.8, we see a similar pattern as when we had $t = 0.25$ where each strategy has a favorable competitive ratio and then the strategy that uses the smallest s value has the largest increase for its competitive ratio and then decreases where the other strategies increase as well and then still increases but at a slower rate towards the end. Also, similar to the previous example, the strategy with the largest slope remains minimal compared to the other strategies until x_m is reached, then the strategy becomes slightly larger than all the other strategies. However, if the requests do not arrive at x_m or after, then the highest s value strategy has the minimal competitive ratio.

34.5 Conclusions

In this research, we made an analysis on a continuous state machine using a set of piece-wise linear strategies and obtained a wide variety of competitive ratios. For every strategy we saw that the competitive ratios were all close to 1 for the first half of the idle duration and had an increase near the halfway point of the idle duration to the end of the idle duration. We ran simulations where we fixed the s value and iterated through a set of t values, when s was arbitrarily small then increasing the duration t increased the overall competitive ratio, the largest peak value was computed, however for that strategy, at around time x_m it had the smallest competitive ratio. When the fixed s value was larger and we iterate through the t values we obtain rather similar competitive ratios for all strategies with the exception of the larger t having a larger peak value midway through the idle period.

When we had a fixed t and we chose a different s value for each strategy, the strategy with the higher s value had a

better competitive ratio, throughout the idle duration, however when the idle time was close to x_m then the highest s had a slightly larger competitive ratio but it had the best competitive ratio through most of the idle period. We see the same behavior when a larger t value is used. Thus with this piece-wise strategies that were developed, the favorable competitive ratios, either the smallest peak value and/or the smallest competitive ratio throughout the idle period was achieved when the duration t and slope s are both arbitrarily high. The worst competitive ratio was when the slope s is low and the duration t is high.

Future work would include analysis of different types of online strategy curves along with different power systems that use different $a(r)$ and $d(r)$ idle and power up cost curves. More analysis with other continuous strategies or more piece-wise function strategies could also be applied on power systems. A randomized approach could also be applied on various power systems.

References

1. Y. Agarwal, S. Hodges, R. Chandra, J. Scott, P. Bahl, R. Gupta, Somniloquy: augmenting network interfaces to reduce pc energy usage. in *Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation, NSDI'09* (USENIX Association, Berkeley, CA, 2009), pp. 365–380
2. Y. Agarwal, S. Savage, R. Gupta, Sleepserver: a software-only approach for reducing the energy consumption of PCS within enterprise environments, in *Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference, USENIXATC'10* (USENIX Association, Berkeley, CA, 2010), pp. 22–22
3. S. Albers, Energy-efficient algorithms. *Commun. ACM* **53**, 86–96 (2010)
4. J. Andro-Vasko, W. Bein, Online competitive control of power-down systems with adaptation, in *Conference on information technology-new generations (ITNG)* (2019), pp. 543–549
5. J. Andro-Vasko, W. Bein, D. Nyknahad, H. Ito, Evaluation of online power-down algorithms, in *Proceedings of the 12th International Conference on Information Technology - New Generations* (IEEE Conference Publications, New York, 2015), pp. 473–478
6. J. Andro-Vasko, W. Bein, H. Ito, G. Pathak, A heuristic for state power Down systems with few states, in *Information Technology - New Generations. Advances in Intelligent Systems and Computing* (2018), pp. 877–882
7. J. Andro-Vasko, S. Avasarala, W. Bein, Continuous state power-down systems for renewable energy management. *Adv. Intell. Syst. Comput.* **738**, 701–707 (2018)
8. J. Andro-Vasko, W. Bein, B. Cisneros, J. Domantay, Online competitive schemes for linear power-down systems, in *17th International Conference on Information Technology–New Generations (ITNG)* (2020), pp. 579–584
9. J. Augustine, S. Irani, C. Swamy, Optimal power-down strategies, in *IEEE Symposium on Foundations of Computer Science* (Cambridge University Press, Cambridge, 2004), pp. 530–539
10. L.A. Barroso, The price of performance. *ACM Queue* **3**, 48–53 (2005)
11. W. Bein, B.B. Madan, D. Bein, D. Nyknahad, Algorithmic approaches for a dependable smart grid, in *Information Technology: New Generations: 13th International Conference on Information Technology*, ed. by S. Latifi (Springer International Publishing, New York, 2016), pp. 677–687
12. A. Kansal, J. Padhye, J. Padhye, J. Reich, M. Goraczko, Sleepless in seattle no longer. Technical report, March 2010
13. S. Nedeveschi, J. Chandrashekar, J. Liu, B. Nordman, S. Ratnasamy, N. Taft, Skilled in the art of being idle: reducing energy waste in networked systems, in *Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation, NSDI'09*, Berkeley, CA (USENIX Association, Berkeley, CA, 2009), pp. 381–394

Ryan Michaels, Md Khorrom Khan, and Renée Bryce

Abstract

Mobile applications are event driven systems that are often driven primarily by user interactions through a GUI. The large event space for mobile applications poses challenges for testing. This work considers the architecture of modern mobile applications to generate test cases that systematically incorporate activities, elements, and events in different sequences for testing. In particular, this work optimizes the presence of different combinatorial-based element sequences of size $t = 2$ inside test suites. The test suites demonstrate increased exploration depth when compared to randomly generated tests, interacting with a wide range of elements and increase code coverage by 0.29–6.79%.

Keywords

Android testing · Mobile application testing · Software engineering · Software testing · Test suite generation

35.1 Introduction

Mobile devices are a fixture in society. In recent surveys, users report being on their mobile devices 3+ hours a day with 90% of their active device time spent on apps [1]. The market is flooded with apps for different purposes and competition

for downloads is often high. If an app crashes on its first launch, a majority of users refuse to use the app again [1]. This rapid growth, heavy usage, and demanding user base raises the need for efficient testing tools.

This paper introduces a tool to generate tests for Android applications using combinatorial-based sequence criteria with onscreen elements. We expand on previous work in which we generated random-walk test suites and then reduced the test suites using combinatorial-based sequence criteria [2]. In the previous work, test suite size was reduced by up to 72.67% while losing at most 0.87% code coverage. The successful application of the criteria for test suite reduction shows a correlation with code coverage and motivates our approach in this work to use the criteria in test generation.

Utilizing the random walk tool created for the previous work, we modify the tool to use a greedy approach that maximizes the presence of element sequences for each criterion in the generated test suite. We generate one test at a time and store the cumulative number of sequences. At each step, we select the element that adds the most new sequences to the set of previously uncovered sequences. If there is a tie between two or more elements, we select one at random. Section 35.3 presents the full definition of the test generation algorithm. We use the same algorithm for all three test generation techniques with differences in the method that selects the “best” next move and one that counts the respective sequences. This paper makes the following contributions:

- A new test generation algorithm that maximize the presence of novel element sequences within test suites.
- Empirical studies that evaluate the code coverage provided by test suites created from our new algorithms, and comparisons between the three algorithms and random walk generation.

R. Michaels (✉)
Department of Computer Science, St Edward’s University, Austin, TX, USA
e-mail: rmichael@stedwards.edu

Md. K. Khan · R. Bryce
Computer Science & Engineering, University of North Texas, Denton, TX, USA
e-mail: mdkhorromkhan@my.unt.edu; renee.bryce@unt.edu

In the remainder of this paper, Sect. 35.2 discusses the current state of test generation in Android and test suite prioritization. Section 35.3 introduces the test generation algorithm. Section 35.4 outlines our experimental setup. Section 35.5 discusses the results. Section 35.6 discusses threats to validity. Section 35.7 provides our conclusions.

35.2 Background

In the realm of Android testing, much of the current literature focuses on test suite generation [3–5]. Current generation techniques fall into one of three categories: user-driven interaction, offline/static navigation of a model of the application, and online/dynamic model navigation [3, 4]. Analysis of test suites primarily focuses on code coverage or faults detected.

Monkey [6] is included in Android development kit and uses a random strategy to generate test cases and is often used as a baseline for comparison in studies. Monkey does not interact with specific elements, but clicks on events at specific (x,y) coordinates on the screen [6]. Monkey “tests” are replayable by providing the same seed to the tool at launch.

Several tools use systematic strategies to generate test cases. MobiGUITAR applies static and dynamic analysis of a GUI model of an application and traverses the model to generate test suites [7]. It follows the same base steps as the Android Ripper [8], but relies on analysis and exploration of the extracted model as opposed to learning based on live exploration of the model [7]. The Android Ripper family expands with the tool AGRippin, utilizing both genetic and hill climbing search algorithms to explore the models to create test cases [9]. Much like other members of the Android Ripper family, a GUI tree is first constructed [8]. In AGRippin the tree is constructed with a hill climbing algorithm. The exploration and translation to test cases is then handled by genetic algorithms as they traverse the GUI tree. An important distinction between other offline search based approaches is that the GUI tree updates when new GUI states are found during exploration and testing via the genetic algorithms [9].

Jaaskelainen et al. and Nieminen et al. create Tema, an on-line testing framework for GUI testing of mobile applications [10, 11]. Tema creates abstract tests based on user informed models of application behavior. These models of abstract actions are independent of any device platform. They identify abstract user actions and the changes they bring about inside the application state. These abstract action sets are translated to traditional test formats via mapping of the abstract user actions to device and application specific actions.

Choi et al. investigate generation-based test suite reduction. They create test suites compatible with regression testing [12]. They focus on coverage of two criterion: abstract

states of the application and the coverage of all loops between them. In a large test suite there are many states and interactions that are covered by dozens or even hundred of test cases. By abstracting states, DetReduce then uses coverage of those abstract states to guide reduction [12]. They further combine test cases that cover the same state with different paths to or exiting from the state [12]. This increases test efficiency by reducing the number of times the phone and application are reset. Like our implementation, they focus on reducing the size of a test suite. However, the interactions they focus on are based on abstract states and loop coverage, where we focus on coverage of specific element sequences.

In their recent work Ratliff et al. explored the possibility of using event sequences to detect vulnerabilities in Android applications [13]. They analyzed hundreds of vulnerability reports to determine which could be attributed to some series of t-events. Like our research, they focused on analysis of open-source applications. A total of 49 vulnerabilities were found to be caused directly via sequences of events, and their work shows that all could be found by at most 5-way event coverage. As part of their research a new tool was developed to measure event sequence coverage of a test suite, and provides not only the coverage provided by the suite, but all possible sequences that are missed as well. Our work differs as we are concerned with the code coverage of the applications under test and we consider sequences of elements.

A new testing framework using IFML has been created by Pan et al. and performs favorable when compared with the tools mentioned above [14]. The goal of IFML is “expressing the content, user interaction and control behaviour of the front-end of software applications”, and has shown success in desktop GUI applications [15]. Their tool Adamant provides a front end for generation of an extended IFML (E-IFML) model and back-end Android test generation. Much like the above AndroidRipper family, Adamant creates a model of the application and then traverses it in a depth first manner. The enhanced modeling capacity that IFML offers allows for superior exploration when compared to the traditional traversal model of Android Ripper [14].

35.3 Test Generation Algorithm

Let us begin by defining actions, elements, and events in the context of our work:

- **Action:** The individual user action such as a screen press, swipe, or text entry, as well as any parameter values associated with the action.
- **Element:** The specific GUI widget that the Action is executed on such as buttons, text fields, and images.
- **Event:** An event is a 2-tuple (Action, Element).

As an example event, a user wishes to add a new email account to their phone. They click “Add Account”. The action performs is a “Click” operation, the element is the “Add Account” button, and the event becomes (“Click”, “Add Account”). In the resulting screen, they fill in the email address via a “Text Entry” action in the “Email” text field. A similar process occurs for the password, and the adding of a new email concludes with the click on “Submit Account” button. Additional sample events include scrolling through the display of unread emails, pressing the phone’s back or home button, opening the support menu, and selecting and unselecting checkboxes in an application’s sound settings.

The sequence criteria consider both inter- and intra-window event combinations. That is, we count events that are invoked on different windows or (Android) activities as well as those which occur on the same window or activity. To use the above example, a user may add an email account through the phone’s setting page, and then scroll through the list of all unread emails. The combinations of events on these two different windows are inter-window interactions, while intra-window interactions are those occurring on the same window. Using our previous example, the process of adding a new account and the process of scrolling through and then reading unread emails are both intra-window interactions.

For this work we examine three sequence criteria, continuing our previous work’s strategy of including element sequences as well as events [2, 16].

- **t-way combinatorial coverage (CCov)** counts every element or event combination without respect to order. Elements and Events do **not** need to be adjacent.
- **t-way sequence-based combinatorial coverage (SCov)** counts element or event combinations with respect to their order of occurrence. This criterion does not require that the events be adjacent to each other.
- **t-way consecutive-sequence combinatorial coverage (CSCov)** counts element or event combinations that appear adjacent to each other in a test case and respects the order of events.

Bryce and Memon conducted empirical studies of GUI applications and noted that their test suites covered less than 50% of all possible two way interactions [17].

This work seeks to generate test cases with higher coverage of two-way element sequences, given their strength in reducing mobile test suites while maintaining code coverage [2]. We refactor the test generation tool of our previous work to evaluate the sequences currently present inside the test suite, and which event execution will add the greatest number of new sequences to the test suite [2]. The Appium server

facilitates the sending of events and confirmation of their execution between the generation platform and the emulator [2, 18]. Experiments are run on a Linux machine running Ubuntu 16.04.

Each sequence criterion guides test generation in a greedy approach to cover two way element sequences. Algorithm 1 illustrates the element selection process. Algorithm 2 illustrates the full generation process. The criterion around which generation is guided is passed in as input. To maximize the presence of unique sequences, the tool counts the previously uncovered sequences and stores the sequences with a hash map.

The initial step of event selection is the same as the previous approach defined in [2]: the tool queries the Appium server for all elements currently on screen. For each element e^i on the screen, the tool generates the set of all new element sequences that e^i would add to the test suite. The tool generates this sequence set via comparison with the set of currently covered sequences. The tool selects the element with the highest number of unknown/non-discovered sequences, and a legal action is randomly selected for execution. If two or more elements provide the same number of new sequences to the test suite, the tie is broken at random. Prior to event execution, the tool updates the current set of sequences with the new discoveries. The emulator then executes the event with the results stored in a JSON file.

After each event executes, the tool checks the exit condition. As in the previous generation tool, a value of 5% is selected as a press of the emulator’s home button or back button [2].

```

Input: XML representation REP, Sequence type T, Current
sequences seq[], Current test case tc[]
Output: Element adding the most coverage
1 current_best = [];
2 count = -1;
3 for Element  $e$  in REP do
4     temp_count = 0;
5     temp_count = T_Count(seq, tc, e);
6     if temp_count > count then
7         count = temp_count;
8         current_best = [e];
9     end
10    else if temp_count == count then
11        current_best.push(e);
12    end
13 end
14 if size(current_best) == 1 then
15     best_element = current_best[0];
16 else
17     i = random_int(0, size(current_best));
18     best_element = current_best[i];
19 end
20 return best_element

```

Algorithm 1: Select_Element

Input: Android Application, Guiding Sequence Criterion T, Number of Tests N

Output: Test Suite designed to maximize presence of input criterion

```

1 num_tests = 0;
2 covered_sequences, test_suite = [];
3 while num_tests < N do
4   open application
5   new_case = []
6   while still in application do
7     curr_elems = All Elements currently on screen
8     to_choose = Select_Element_T(curr_elems,
9     covered_sequences, new_case);
10    action = Pick_Action(to_choose)
11    emulator.execute(to_choose, action)
12    mark_element_SCov(to_choose, covered_sequences,
13    new_case);
14    new_case.append([to_choose, action])
15    exit = random(0,1)
16    if exit < 0.05 then
17      | Exit application
18    end
19  end
20  test_suite.append(new_case)
21  num_tests++;
22 end
23 Return test_suite

```

Algorithm 2: Test suite generation algorithm

35.4 Experimental Setup

In this paper we examine the following research questions:

1. **RQ1: What is the exploration depth of on-screen elements in test suites generated with CCov, SCov, and CSCov?**
2. **RQ2: What is the exploration depth of sequences in test suites generated by CCov, SCov, and CSCov?**
3. **RQ3: What level of code coverage is achieved by test suites generated with CCov, SCov, and CSCov?**

Subject Applications This paper analyzes test suites for a set of open source Android applications from F-Droid [19]: Hourly Reminder [20], Memento [21], Poet Assistant [22], and Pointeuse [23]. These applications run on Android 19 (version 4.4). Direct access to the application source code is required to instrument the apps to create code coverage reports with through a JaCoCo plugin [24, 25]. JaCoCo is a free open-source code coverage tool for Java. No other modification were made to the application code. The applications have between 3 to 10 activities and are between 5873 to 60,222 lines of code (LoC) as shown in Table 35.1.

Test Generation Test suites are generated for each application according to the algorithms described above using the combinatorial-based sequence criteria. The tool generates three hundred test cases per application per criterion. This

Table 35.1 Application details

Criterion	Hourly reminder	Memento	Poet assistant	Pointeuse
LoC	60,222	9082	23,963	5873
Branches	4605	457	1993	272
Methods	2202	470	1301	146
Classes	465	141	197	35
Activities	3	5	10	6

matches the size of the test suites produced for the experiments of the previous work [2].

Sequence Criteria Counts of element and event sequence criteria are extracted from each test case. For each of the novel combinatorial criteria, scripts generate the exhaustive set of each criterion that the full test suite covers. We store the sequences that individual test cases cover and the totality of sequences that the test suite covers. Tables 35.3, 35.4, and 35.5 present information on the total elements, events, and sequences covered for each application by the generation algorithm. The results demonstrate increased growth in discovery of all criteria, especially for *SCov* and *CSCov*, with increases ranging from 5% to 88%.

Code Coverage JaCoco extracts code coverage data from each individual test case [24, 25]. When the test suite generation completes, scripts calculate the totality of the test suite's code coverage using JaCoCo.

Evaluation Metrics This paper evaluates the test suites by two metrics: (1) the overall number of sequences for each novel combinatorial criterion and (2) the code coverage. We compare the results to an algorithm with a random walk strategy for a baseline comparison.

35.5 Results

The results demonstrate variation among the criteria and apps. We summarize the results by research question.

RQ1: What is the exploration depth of on-screen elements in test suites generated with CCov, SCov, and CSCov?

Tables 35.2, 35.3, 35.4, and 35.5 display the number of elements, events, and sequences covered in each experiment. In all but one case, Memento *CSCov generation*, the test suites generated by the optimized algorithms cover more elements than those generated by random. This behavior matches expectations, as the algorithms prioritize covering a wider set of element sequences during the generation process. This increase in coverage is not carried over to the number of events for every algorithm. *CCov generation* across the board

Table 35.2 Application statistics: random walk

Criterion	Hourly reminder	Memento	Poet assistant	Pointeuse
Elements	165	79	107	112
Events	212	103	141	225
Element_CCov	7067	3098	2607	4711
Element_SCov	7430	3419	2833	4951
Element_CSCov	1505	575	648	1289
Event_CCov	9310	4531	3817	11,225
Event_SCov	9795	4942	4132	11,798
Event_CSCov	1749	811	844	2225

Table 35.3 Application statistics: CCov generation

Criterion	Hourly reminder	Memento	Poet assistant	Pointeuse
Elements	183	81	125	124
Events	205	96	147	207
Element_CCov	5663	2062	2317	4656
Element_SCov	5778	2246	2548	4921
Element_CSCov	820	328	509	1012
Event_CCov	6210	2548	2917	9519
Event_SCov	6328	2805	3141	10,011
Event_CSCov	903	377	593	1808

Table 35.4 Application statistics: SCov generation

Criterion	Hourly reminder	Memento	Poet assistant	Pointeuse
Elements	191	90	139	129
Events	232	113	173	230
Element_CCov	13,301	4769	4028	6459
Element_SCov	13,856	5211	4304	6740
Element_CSCov	1786	753	767	1353
Event_CCov	14,941	6689	5162	11,499
Event_SCov	15,540	7277	5474	12,048
Event_CSCov	1903	950	907	1898

covers fewer unique events than random. *CSCov generation* covers more events in Memento and Poet Assistant, but fewer in Hourly Reminder and Pointeuse. *SCov generation* covers more events in every instance.

RQ2: What is the exploration depth of sequences in test suites generated by CCov, SCov, and CSCov?

For coverage of sequences, a similar pattern emerges. *CCov generation* covers fewer sequences of every type when compared to random generation for all applications, dramatically so in event sequences. *CSCov generation* generates more element and event sequences than random for Hourly Reminder, Memento, and Poet Assistant, while producing more element sequences for Pointeuse. Random generation covers more event sequences for all criteria in Pointeuse. *SCov generations* generates more sequences for every application and criterion except for *Event CSCov* in the Pointeuse application when compared to random. These sequence statistics are displayed in Tables 35.2, 35.3, 35.4, and 35.5.

Table 35.5 Application statistics: CSCov generation

Criterion	Hourly reminder	Memento	Poet assistant	Pointeuse
Elements	184	78	120	125
Events	213	95	145	190
Element_CCov	9557	3611	3117	5611
Element_SCov	10,204	3858	3436	5831
Element_CSCov	2411	771	845	1460
Event_CCov	11,304	4904	4301	10,197
Event_SCov	12,052	5227	4639	10,669
Event_CSCov	2595	940	999	2013

Table 35.6 Application code coverage by strategy

Strategy	Hourly reminder	Memento	Poet assistant	Pointeuse
Random walk	62.71%	78.34%	63.17%	68.82%
CCov Gen	63.02%	75.61%	61.23%	72.11%
SCov Gen	64.00%	78.63%	61.59%	74.71%
CSCov Gen	64.28%	77.81%	64.53%	75.61%

CCov generation's holding the worst performance is expected. It covers more elements than random generation, but finds fewer events and sequences of all types. The properties of the *CCov* criterion may hinder exploration. For example, clicking “New Note” and then entering text in “Text Entry” is considered the same as entering text into “Text Entry” and later clicking on “New Note”. *SCov* treats those two instances as distinct sequences. Generation guided by *SCov* would consider the second sequence as a novel one to be added, as opposed to *CCov* which sees the second as being already provided. Considering order of occurrence can greatly expand the potential sequence space. With this greater space *SCov generation* covers more sequences than the random strategy. Test suites guided by *SCov* cover 5–88% more element sequences than random generation. When compared to *CCov*, *SCov generation* discovers more sequences of every type, with a minimal increase of 37% for element sequences and 5% for event sequence. For several criteria, including *SCov* sequences in Memento and Hourly Reminder, *SCov generation* doubles the sequences found by *CCov generation*. *CSCov generation* also outperforms random generation. *CSCov generation* produces 13–60% more element sequences than random. It significantly outperforms *SCov generation* on the number of *CSCov* sequences. *CSCov generation* covers fewer unique *CCov* and *SCov* sequences when compared to *SCov generation*.

RQ3: What level of code coverage is achieved by test suites generated with CCov, SCov, and CSCov?

The code coverage provided by the experiments is shown in Table 35.6. For all applications *CCov generation* produces the lowest levels of code coverage out of the three strategies. *CSCov generation* test suites produced the highest code coverage in three of the four applications, with 0.9–2.95% higher coverage compared to *SCov*. This improvement stems from

the more stringent selection criterion that considers only the previous element in the test case as opposed to all elements visited in the current suite. It is clear that the order of the interaction matters, given the disparity in both code coverage and total sequences generated between *SCov* and *CCov*. This is illuminated via a simple example of a calculator, the sequence 2/3 produces a different result than the sequence 3/2. *CCov* would treat those sequences as the same while *SCov* and *CSCov* detect them as different sequences. A further example from the Memento app: a user enters text and then saves the note; a user saves the current note then enters text. *CCov* sequences treats the two interactions as the same. *SCov* would cover each event as independent of one another, as would *CSCov* if the events occurred consecutively.

The new generation techniques performed well in comparison to random. *CCov* generation outperforms random for Pointeuse and Hourly Reminder in code coverage, providing 3.29% and 0.31% more coverage, while providing 1.99% and 1.94% less code coverage in Memento and Poet Assistant. *SCov* generation outperforms random in all but Poet Assistant, providing 0.29%, 5.89%, and 1.29% increased coverage, while losing 1.59% coverage in the Poet Assistant app. *CSCov* generation outperforms on all but Memento, especially in Pointeuse and Hourly Reminder. It provides between 1.36–6.79% more coverage while losing 0.53% coverage in Memento. Through increasing the coverage of the sequence criteria, in 6 of 8 cases the test suites generated by our algorithms achieve superior performance to the random in code coverage. In the instances where they do not increase sequence coverage, the exploration provided by the generation algorithm allowed for deeper exploration of small interactions, increasing the code coverage provided in the Pointeuse app.

The two test suites where the new generation techniques failed to outperform random occurred in the applications with the lowest total number of sequences. When compared against Hourly Reminder and Pointuisse, which contain up to two and a half times more sequences as Memento and Poet Assistant, the algorithms presented in this paper outperform random quite significantly, with even the worst performing criterion offering higher. This suggests that applications with a larger sequence space (containing more on-screen elements or a wider set of allowable actions) will benefit from generation via *SCov* and *CSCov* when compared to traditional strategies.

35.6 Threats to Validity

There are several threats to validity in this paper that prevent us from generalizing the results to all applications. We create test suites for four applications. Studies using other applications with different characteristics may produce different

results. We minimize this threat by using applications of differing sizes and purposes. The algorithms do not consider advanced tie breaking procedures to select the “best” out of tied elements. Different tie breaking strategies may produce different results.

35.7 Conclusions

This paper presents an algorithm for Android Test Suite Generation using *Combinatorial Coverage (CCov)*, *Sequence-based Combinatorial Coverage (SCov)*, and *Consecutive-sequence Combinatorial Coverage (CSCov)* as the guiding criteria. This algorithm maximizes the presence of the criterion, focusing on elements sequences of size $t = 2$ with the goal of increasing code coverage via increased sequence coverage. Test suites for each criterion contain more element interactions than test suites generated with *Random*. Test suites generated by *CSCov* achieved the highest code coverage for three out of four applications, beaten by *SCov* in the application Memento. Both *CSCov* and *SCov* cover more code than *Random* test generation for three applications. Both cover more element sequences of all types and more event sequences for all but Pointeuse. *CCov* generation achieves low code coverage and sequence coverage for all three algorithms. It outperformed random in Hourly Reminder and Pointeuse for code coverage, but did not improve on total sequence coverage. This weakness comes from *CCov* strategies treating sequences such as (“New Note”, “Text Entry”) as identical to (“Text Entry”, “New Note”). The results suggest that applications with a deep sequence space will benefit from *SCov* and *CSCov* generation. Future work will explore optimization strategies focused on tie-breaking procedures to maximize exploration.

References

1. D. Chaffey, Mobile marketing statistics compilation (2020). <https://www.smartinsights.com/archive/mobile-marketing/mobile-marketing-analytics/>. Accessed 10 Nov 2020
2. R. Michaels, D. Adamo, R. Bryce, in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)* (IEEE, New York, 2020), pp. 0598–0605
3. S.R. Choudhary, A. Gorla, A. Orso, in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (IEEE, New York, 2015), pp. 429–440
4. D. Amalfitano, N. Amatucci, A.R. Fasolino, P. Tramontana, in *2015 30th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW)* (IEEE, New York, 2015), pp. 50–57
5. D. Amalfitano, N. Amatucci, A.M. Memon, P. Tramontana, A.R. Fasolino, *J. Syst. Softw.* **125**, 322 (2017)
6. Google Inc. Ui/application exerciser monkey. <http://developer.android.com/tools/help/monkey.html>. Accessed 10 Nov 2020
7. D. Amalfitano, A.R. Fasolino, P. Tramontana, B.D. Ta, A.M. Memon, *IEEE Softw.* **32**(5), 53 (2014)

8. D. Amalfitano, A.R. Fasolino, P. Tramontana, S. De Carmine, A.M. Memon, in *2012 Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering* (IEEE, New York, 2012), pp. 258–261
9. D. Amalfitano, N. Amatucci, A.R. Fasolino, P. Tramontana, in *Proceedings of the 3rd International Workshop on Software Development Lifecycle for Mobile* (2015), pp. 5–12
10. A. Jaaskelainen, M. Katara, A. Kervinen, M. Maunumaa, T. Paakkonen, T. Takala, H. Virtanen, in *2009 31st International Conference on Software Engineering-Companion Volume* (IEEE, New York, 2009), pp. 112–122
11. A. Nieminen, A. Jaaskelainen, H. Virtanen, M. Katara, in *2011 11th International Conference on Quality Software* (IEEE, New York, 2011), pp. 131–140
12. W. Choi, K. Sen, G. Necul, W. Wang, in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)* (IEEE, New York, 2018), pp. 445–455
13. Z.B. Ratliff, D.R. Kuhn, D.J. Ragsdale, in *2019 IEEE 19th International Conference on Software Quality, Reliability and Security (QRS)* (IEEE, New York, 2019), pp. 159–166
14. M. Pan, Y. Lu, Y. Pei, T. Zhang, J. Zhai, X. Li, *J. Syst. Softw.* **159**, 110433 (2020)
15. IFML, The interaction flow modeling language. <https://www.ifml.org/>. Accessed 10 Nov 2020
16. Q. Mayo, R. Michaels, R. Bryce, in *2014 IEEE Seventh International Conference on Software Testing, Verification and Validation Workshops* (IEEE, New York, 2014), pp. 128–132
17. R.C. Bryce, A.M. Memon, in *Workshop on Domain Specific Approaches to Software Test Automation: In Conjunction with the 6th ESEC/FSE Joint Meeting* (2007), pp. 1–7
18. Appium: automation for apps. <http://appium.io/>. Accessed 10 Nov 2020
19. F-droid - free and open source android app repository. <https://www.f-droid.org/>. Accessed 10 Nov 2020
20. Axet, Hourly reminder. android hourly reminder application. <https://gitlab.com/axet/android-hourly-reminder>. Accessed 10 Nov 2020
21. N. Fathollahzade, Memento: a simple note taking app for android. <https://github.com/yaa110/Memento>. Accessed 10 Nov 2020
22. C. Alvarez, Poet assistant: android app with rhyming dictionary, thesaurus, and dictionary, with text-to-speech functionality to read your poem. <https://github.com/caarmen/poet-assistant/>. Accessed 10 Nov 2020
23. Landais, Pointeuse: application android qui offre les fonctionnalités d'une pointeuse. https://github.com/alandais/Pointeuse_S3I. Accessed 10 Nov 2020
24. EclEmma, Jacoco - jacoco java code coverage library. <https://www.eclemma.org/jacoco/index.html>. Accessed 10 Nov 2020
25. Gradle Inc., The jacoco plugin. https://docs.gradle.org/current/userguide/jacoco_plugin.html. Accessed 10 Nov 2020

Jonathan de Oliveira T. Souza, Adler Diniz de Souza,
Leandro G. Vasconcelos, and Laercio A. Baldochi

Abstract

In software engineering, bad smells are symptoms that something may be wrong in the system design or code. The term bad smell usually describes a potential problem with known consequences, and also known solutions. In the case of usability, a bad smell is a hint of poor interface design, that makes it difficult for the end-user to accomplish common tasks. This paper presents the most relevant usability smells reported in the literature, as well as tools, methods and techniques that support the process of detecting usability anomalies.

Keywords

Usability smells · Bad usability · Bad usability smells · Catalog of bad smells · Web applications · Mobile applications · Automatic detection tools · Usability problems · Detection of user behavior · User satisfaction

36.1 Introduction

In the last decade, web applications have become part of our daily lives. Common activities such as shopping, home-banking, reading news and social interactions depend on these applications. Despite its massive use, many of these

J. de Oliveira T. Souza
POSCOMP, Federal University of Itajubá, Itajubá, Brazil
e-mail: otsjonathan@unifei.edu.br

A. D. de Souza · L. A. Baldochi (✉)
Institute of Mathematics and Computing, Federal University of Itajubá,
Itajubá, Brazil
e-mail: adlerdiniz@unifei.edu.br; baldochi@unifei.edu.br

L. G. Vasconcelos
National Institute for Space Research, São José dos Campos, Brazil

applications present usability problems that hinder their use [7].

In order to minimize usability problems commonly found in web applications, several techniques for usability evaluation have been proposed [3], among them, an example is the SUS method developed by John Brooker, which consists of a questionnaire with well-formulated questions [2].

According to Ribeiro et al. [14], “the success of a web application is highly related to its interface usability”. Weichbroth [15] reinforces this statement saying that software systems must provide a pleasant user experience for users to complete a task. Another important aspect to consider is the fact that web applications are constantly evolving, thus regular usability assessments are required to ensure a good level of usability and user satisfaction [1].

Recent research in usability assessment has exploited the concept of smells in order to identify usability issues. This concept was initially introduced by Fowler in 1999, but in the context of code design, where it was identified that possible anomalies could be presented in application source code, which hindered the development, interpretation and maintenance work, coining the term “code smells” [4]. Following, Hermans et al. conducted a study relating the catalog of smells created by Fowler in the context of spreadsheets, in order to identify possible anomalies in this environment [10].

The literature reports recent work focusing on user behavior that aims to reproduce actions performed by users on the interfaces of websites to determine weaknesses in web/mobile applications. In this context, some techniques and tools were proposed, as, for example, in [1, 6–9, 12, 13]. The search for patterns and usage sequences that can represent a usability problem, regardless of the tool or approach employed, contributed to compose a catalog of bad usability smells.

The purpose of this review is to present the usability smells concept, based on work already done, and to show a catalog of smells and applied tools/techniques to identify usability

problems. It is worth noting that the study of usability smells aims not only to contribute to the detection of the weaknesses of an application, but also to suggest refactorings so that the evaluators and those involved can guarantee the good usability of the system. However, it is not the focus to present the refactoring suggestions proposed by the related researchers.

This paper is organized as follow. Section 36.2 presents the research methodology employed to produce this review. Section 36.3 presents the results of the literature review and an analysis of the extracted data, that is, answering the defined research questions. Section 36.4 presents a discussion of the results. Finally, Sect. 36.5 presents the conclusions.

36.2 Research Methodology

To present the knowledge of Usability Smells, we performed the planning and reporting steps of the systematic literature review method. The planning stage is presented in the following subsections, while the reporting stage comprises the content presented in Sect. 36.3 of this document.

36.2.1 Planning

In this stage, the following activities were carried out, respectively: (1) definition of research questions, (2) development of the search strategy, (3) selection of primary studies, (4) quality assessment, and (5) development of the data extraction strategy. These steps are presented in detail below.

1. *Research questions:* To maintain the focus of the study, it is essential that research questions are defined, in order to limit the subject and prevent the researcher from distancing himself from the research objective [16]. In order to achieve the objective of this study, which is the presentation of the concept of Usability Smells, five guiding research questions (RQ) were proposed, accompanied by their motivations (M):

RQ1: How is the concept of Usability Smells defined?

M1: Discover the definition of Usability Smells, understand its application and its importance.

RQ2: Which smells are reported in the literature?

M2: Find out what anomalies can be identified in the web and mobile contexts. Find and compile the smells catalogs already registered.

RQ3: Which tools, techniques or methods allow the identification of bad usability smells in interactive applications?

M3: Find out which tools, techniques or methods exist to find usability problems, and also understand how this process is carried out.

RQ4: In what contexts were these tools, techniques or methods applied?

M4: Understand in which contexts these instruments can assist in the detection of usability problems.

RQ5: What is the number of anomalies identified by the tools, methods or techniques?

M5: Discover the detection potential of a tool, method or technique.

2. *Search strategy:* The search strategy is directly related to the data source, where the information we want to find is found, and also to the way we search for the information that can contribute to answer the research questions defined in the previous activity.

The main digital libraries that were used to find the primary studies were: Scopus, ACM and IEEEExplore. The snowballing technique, which consists of searching for new sources of data through the references of a document, was also applied to find more evidence [16].

To perform the search process, a search string was proposed and used in each digital library. During the development of the search string, the concepts of usability, tool and web application were adopted, and then keywords and synonyms were obtained. The defined search string was:

(“Usability Smells” OR “Bad Usability”) AND (“Detect” OR “Tool” OR “Refactor”) AND (“Web” OR “Web Development” OR “Web Application”)

The search procedure and the criteria for selecting primary studies are described in detail in the next subsection.

3. *Selection of primary studies:* After executing the search string, the documents returned were initially evaluated by reading the title, abstract, keywords and conclusion. Some criteria for inclusion or exclusion of studies were established, in order to provide greater precision for the choice of primary studies.

The inclusion (IC) and exclusion (EC) criteria are shown in Table 36.1.

It is essential to consider documents obtained from the “snowballing” process, which can be relevant or complementary documents on the subject. For example, the studies presented in [3, 9, 11].

4. *Quality assessment:* A questionnaire-type instrument, in the 5-point format on the Likert scale, was designed to score and detect the level of contribution that a selected article has to the research carried out. The questionnaire

Table 36.1 Inclusion criteria

Identifier	Description
IC-1	The document must be an article, magazine, book or conference proceedings
IC-2	The document must have been published in English
IC-3	The document must contain content directly related to usability smells
IC-4	The document must present a practice of capture user behavior in web or mobile tools
IC-5	The document must present a catalog of smells when the author presents a tool
IC-6	The document that presents a tool for the detection of usability smells must have at least one case study of its application
IC-7	The document must present a technique for visualizing usability smells
EC-8	Documents that have a keyword in their title, abstract or text, but are not related to the theme
EC-9	Documents that were published before 2014
EC-10	Duplicate documents obtained through 2 or more different virtual repositories

consists of five closed questions, two of which are objective and the rest subjective. The possible answers ranged from “totally disagree” (−2) to “totally agree” (+2), with the answer “neutral” being worth 0 points.

The subjective questions were:

1. Did the study provide details about cataloged smells?
2. Did the study show how to use the proposed tools or techniques?
3. Did the study present results and analyzes them in a clear and satisfactory manner?

The objective questions were:

1. Was the study published in renowned journals or conference proceedings? For this question, the “H5-index” should be used, applying the adaptation of the possible answers to the following format: between 0 and 25 (totally disagree), between 25 and 50 (partially disagree), between 50 and 75 (neutral), between 75 and 100 (partially agree), lastly greater than 100 (totally agree).
2. Was the study cited by other authors? Responses must be represented by: no citations (totally disagree), between 1 and 5 citations (neutral), more than 5 citations (totally agree).

The result of the study evaluation is calculated using the arithmetic mean of the individual points of each question. The result value is not used as an exclusion criterion. However, this value represents a contribution rate and relevance of the document in question to the research.

5. *Data extraction strategy*: This step is important to obtain a classification of the relevance of the selected documents. For this, a checklist of questions was elaborated that portrays some points that we wish to identify in the selected studies. The checklist reinforces the defined research questions.

It is worth mentioning that, in the first moment, after the selection of primary studies, all documents were read, and false positives were discarded, which did not present expected contributions after the selection activity in which they had been approved. Then, another reading was performed with the purpose of extracting the data based on the questions presented in Table 36.2.

After analyzing the documents, only nine of them were identified as relevant as input for analysis and provide answers to contribute to the research. Even after performing the new classification of documents, the level of relevance is not applied as an exclusion factor for any of them.

36.3 Results

Finally, after the first execution of the search string, 84 documents from the selected virtual repositories were obtained, 34 of them from ACM, 46 from Scopus and 4 from IEEEExplore. Then, applying the inclusion and exclusion criteria, a total of 18 documents were returned, 4 of them from ACM, 11 from Scopus and 3 from IEEEExplore. However, there were 4 duplicate files, that is, presented in more than one repository, and therefore the duplicates were eliminated. The final number of documents obtained after the filtering step was 14.

As mentioned, the “snowballing” technique was used, which returned 10 documents that were also evaluated in the same way. Finally, the quality assessment of all texts was carried out and the data that make up this section were extracted.

36.3.1 Usability Smells

The Usability Smells concept was derived from refactorings approach presented by Martin Fowler, which consisted of finding “bad smells”, that is, anomalies that affected the characteristics of interpretation, reusability and maintainability of the source code, mainly in the development phase, in this context these anomalies were called code smells [4].

Table 36.2 Checklist for data extraction

Questions
Does the author present a catalog of smells?
Did the document present a tool, technique or method for identifying smells?
In what context was the tool, technique or method applied?
Did the document present a case study, in which the presented tool was applied?
Did the document present concrete analyzes, that is, information that validates both the test performed and the conclusions drawn?
How effective was the tool, technique or method for the task of detecting usability smells?

So, unlike the code smells scenario, many approaches have emerged to identify usability problems in interfaces applying the concept, in order to complement the assessment methods existing traditional methods, such as questionnaires, think-aloud protocol, analysis of remote logs and tests [3, 15]. Thus, the Usability Smells concept emerged, which can be defined as a strategy for identify the weaknesses of an application and assume some difficulties that eventually a user can develop in a way unexpected. In addition, it is important to understand how the problem occurred and provide suggestions to resolve it, so you need to collect data on the behavior of the user [6, 8, 14].

The task-based usability evaluation has always been classified as long, exhaustive and costly, even including experts on the subject, as problems can be subjective and are associated with software complexity [7, 14]. To provide more quality to web applications, some studies have proposed tools for automatic detection of smells, also aiming to optimize the resources involved in this process. Some examples are: Auto-QUEST (Automatic Quality Engineering of Event-driven Software), USE (UseSkill Extension), GUISurfer, MUSE (Mobile Usability Smell Evaluator), and USF (Usability Smells Finder) [1, 7, 8, 13, 14].

From the selected studies we are able to answer the first research question: **QP1. How is the concept of Usability Smells defined?** It is possible to understand that any user behavior that hinder or prevent the user from performing any task is considered a particular problem of usability. The automation of this process, evidencing the user's behavior (log analysis) and patterns that point out the affected elements provide the opportunity for a better understanding of the case [6, 7, 13]. Therefore, the use of smells as parameters can make evaluations more agile and complete, since the concept extends to providing suggestions for refactorings and helping those with less experience in usability evaluation [14].

36.3.2 Bad Usability Smells Catalog

Recent studies try to understand the user's behavior and define the possible "usability smells" from the catalog of "bad usability smells" established by Fowler [4]. Thus, today we have a large number of anomalies that can be found

in interactive applications. The bad usability smells catalog serve as guidelines for identifying the expected user behavior and finding a usability problem [1].

Fields marked with (*) on Table 36.3 refer to smells found in more than one study, however they may differ in relation to the metrics adopted by the researcher, and also to some conceptual point.

With respect to **RQ2**, the selected studies allowed to collect the various smells proposed by the authors. It is worth mentioning that in each study the author used a tool or technique to validate the identification of the anomaly [1, 6–8, 13, 14]. Thus, we have a representative collection of smells, which was referred to in this document as a general catalog of bad usability smells.

The presentation of the general catalog consists of listing the works and their respective smells, as shown in Table 36.3.

Considering the 7 usability factors proposed by Garrido et al. in [5], a concept map was developed to list the smells in the general catalog. For this, the identifier associated with the smells presented in Table 36.3 were used in order to classify the corresponding usability factor, as shown in Fig. 36.1.

From Fig. 36.1, we can identify that the effectiveness and comprehensibility factors have the largest amount of related smells – both have 11 smells – followed by the factors, in their respective order: navigability, credibility, customization, accessibility and learning. The three most prominent factors have the following meanings defined by Garrido et al. [5]:

Efficiency The application provides more agile paths for advanced or recurring users. The contribution made to this factor was to consider the optimization of some process during the use of the application.

Understandability The organization of the page structure and content provides the user with a better understanding of the purpose of the application and how to use it. No contribution was made to this factor, since its definition appears to be complete and coherent.

Navigability Refers to the quality of navigation within the application, that is, it guarantees easy access to the content. A contribution was made to this factor, there may be an

Table 36.3 General catalog of bad usability smells

Authors	Bad Smells	Identifier	
Harms et al. [8]	Missing feedback* important task	1	
	Required inefficient actions	2	
		3	
	High website element distance	4	
	Laborious task	5	
	Cyclic task	6	
Ribeiro et al. [14]	Lonely action	7	
	Too many layers*	8	
	Undescriptive element*	9	
	Missing feedback*	10	
	Shotgun surgery	11	
	Too many layers*	12	
	Middle man	13	
Almeida et al. [1]	Information overload	14	
	Inappropriate intimacy	15	
	Feature envy	16	
	Too small or close elements	17	
	Too close links	18	
Paternó et al. [13]	Distant content*	19	
	Too small section	20	
	Bad readability	21	
	Long forms	22	
	Undescriptive element*	23	
	Misleading link	24	
	No processing page	25	
	Free input for limited values	26	
	Unformatted input	27	
	Short input	28	
	Unnecessary bulk action	29	
	Grigera et al. [7]	Overlooked content	30
		Distant content*	31
No client validation		32	
Late validation		33	
Abandoned form		34	
Scarce search results		35	
Useless search results		36	
Wrong default value		37	
Unresponsive element		38	

unnecessary dependency between pages to access content, this anomaly was initially presented by Fowler through the smell Middle Man [4].

Finally, the other factors presented complete definitions and no contributions were needed. It is worth mentioning the accessibility factor that has only one smell associated due to the possible problem of using the application by users with vision impairment. The smell Bad Readability can represent this threat, however, it can also represent the understandability factor, since the text size is related to a poor

page structure. Lastly, the learning factor has no associated smell. One possible explanation for this is that anomalies alone cause a learning problem.

36.3.3 Tools and Techniques

This section presents the tools and techniques found in the literature for the task of detecting usability smells. The results presented are the answer to **RQ3**. In general, all tools use some technique to collect user interactions automatically, however they differ in relation to the procedures for analyzing the collected data.

As discussed earlier, each tool was developed with the purpose of identifying only the smells presented in its catalog. It is important to note that, during the data collection process, it is extremely important to guarantee user privacy. Thus, personal information such as bank account data and passwords are not registered [1, 7, 8, 13, 14].

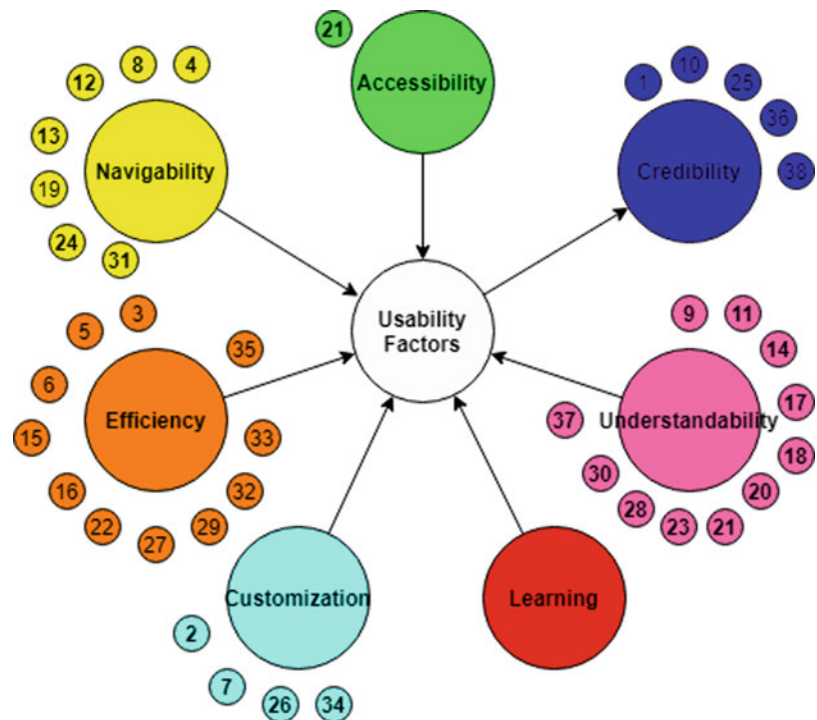
36.3.4 USF – Usability Smells Finder

The USF tool, created by Grigera et al. [6, 7], uses a three-step approach to identify usability problems in web applications. These steps are event logging, usability smells detection and usability smells reporting. The first step is performed on the client side and the other two on the server side [7].

In the first stage, the objective is to collect data on user behavior, using a set of usability events. For this, a script is inserted into the web application. Instead of collecting the log of all user interactions and forwarding directly to the detection stage, the authors suggested a technique to filter the logs, still on the client side, based on events that are strongly related to the existence of some bad smell [7]. For example, the event “attempted clicks” is related to usability smell “unresponsive element”, which can be identified when a user clicks on an element that has no reaction.

In the second stage, the process consists of three steps: classification of events, synthesis of data and evaluation of usability smells. This step takes place on the server side, where the user logs are actually analyzed. For this, the tool has several algorithms, called usability smells finder [6], whose purpose is to detect a specific usability smell. In the event classification step, as soon as a log is received on the server side, it is associated with an appropriate “finder” algorithm, which classifies it regarding the affected element. In the second step, the essential data of the classified event is extracted, and some counter can be updated relating the affected element [7]. Finally, an evaluation of the affected element is conducted to verify the presence of a smell.

Fig. 36.1 Conceptual map of the relationship between smells and usability factors



In the last step, the tool provides a report of the bad smells found in the application together with the identification of the affected element and suggestions for refactorings. Also, some additional information can be provided to the evaluator for a better understanding of the problem, as, for example, in a “scarce search result” problem, in which he can present a ranking with the search words. The USF also offers a real-time view of the bad smells identified [7].

36.3.5 MUSE – Mobile Usability Smell Evaluator

MUSE is a tool developed by Paternó et al. [13] in order to identify usability problems in mobile web applications. The tool is a web proxy, acting as a broker between the end-user and the application. User behavior data is captured using a “JavaScript logger” that is inserted on the website.

MUSE allows the capture of various types of interaction on mobile devices (eg: touch, double touch, drag, pinch, drawing, pressing, rotating, orientation changes), in addition to using a library that also allows the capture of native device data (eg gps, accelerometer) [13]. For data collection, the tool allows users to define tasks that are used to detect bad smells.

Following, the tool performs the analyzes automatically using an algorithm presented in [13], which follows predefined steps looking for patterns that represent some usability anomaly. However, according to Paternó et al. [13], this task

can only be automated if a bad smells dictionary is defined, that is, an XML document that contains data that represents usability problems.

In conclusion, the tool also provides support with a graphical representation of the actions taken by the user through a timeline, which allows a better interaction with the data to understand and formulate concrete considerations regarding usability assessments.

36.3.6 AutoQUEST – Automatic Quality Engineering of Event-driven Software

Harms et al. [9] proposed an approach to track user behavior and actions in web applications, based on DOM (Document Object Model) events that occur on these pages. They developed a tool capable of capturing and recording all actions performed by users on websites. Like the previous tools, the approach requires JavaScript to collect user logs, however it is necessary that the script be present on each page of the application.

To perform a given task, a user must perform several actions, also called task events. In order to reconstruct a user’s actions timeline, the script sends the collected data to the AutoQUEST server, where the analysis is performed [8, 9].

The approach presents the task trees model. It is based on a Trie tree, also known as prefix tree. The task tree provides a temporal relationship to the order of execution of user actions and is composed of sequence and iteration. A

sequence reflects “a task that has one or more children and they are executed in their respective orders” [8]. Whereas an iteration is “a task that has only one child that is executed none or more times” [8]. After tracking the user’s actions on the website, an ordered list of actions is generated, which is submitted to the analysis steps: iteration detection, sequence detection and repetition detection [9]. Thus, it is possible to generate a Trie tree that represents the user’s behavior when performing a task.

To find bad smells the task tree is analyzed in order to find sequences of expected events, and also values that represent the existence of a smell from some related metric. This information is presented in the catalog proposed by Harms et al. [8].

36.4 Discussion

Considering the results obtained, it was possible to answer the research questions RQ1, RQ2 and RQ3 in the previous section. After the presentation of the smells contained in the literature and the tools for the detection of usability smells, it is possible to answer the research questions RQ4 and RQ5.

For RQ4, we have that all the tools are able to work in the web environment, except the GUISurfer tool that operates in the desktop context [1]. Finally, RQ5 can be answered by analyzing Table 36.3, where it is clear that the USF tool developed by Grigera et al. [7] is capable of detecting a large number of anomalies in relation to other tools. However, it is important to note that this issue is not about the efficiency of one tool over the other.

To deal with the comparison between two tools, the following considerations must be made:

- Detect the same anomaly; and,
- Consider the same test cases.

Thus, the desirable results are:

- Task completion time;
- The graphical visualization of the data related to the user’s behavior that were captured by the tool. This result must be analyzed subjectively by a usability expert;
- Possibility of readjustments, learning; and,
- The affected element.

Based on this comparison strategy, we can identify which approach is best for the task of detecting usability smells, and also the tool with the best performance to perform this activity.

36.5 Conclusion

This research was developed in order to understand the state of the art on the subject Usability Smells, presenting approaches for the detection of anomalies in interactive applications. All smells already recorded in the literature were reported, as well as 3 tools developed for the task of detecting usability problems. It could be understood that the development of tools to identify anomalies in an automatic way is more effective than manual tasks, however, there is still no precision on the part of these instruments when claiming a possible problem, so it is still essential for a professional to validate the final results. In addition, the tools proved to be able to present the data that resulted in the identification of an anomaly to the evaluator.

From the developed conceptual map, it is expected that the classification of smells will provide insights in other researchers so that they can make new contributions in relation to the discovery of new smells and new strategies for their identification.

As proposals for future work, there are different topics to investigate, such as: creating metrics to identify bad usability smells; implementation of efficient algorithms for log analysis, in order to detect smells; detection of smells during user browsing; and study of the relationship between bad usability smells and usability heuristics.

References

1. D. Almeida, J.C. Campos, J.a. Saraiva, J.a.C. Silva, Towards a catalog of usability smells, in *Proceedings of the 30th Annual ACM Symposium on Applied Computing, SAC '15*, (New York, 2015), pp. 175–181
2. J. Brooke, Sus: A “quick and dirty” usability, in *Usability Evaluation in Industry*, (Taylor & Francis, London, 1996), p. 189
3. A. Fernandez, E. Insfran, S. Abrahão, Usability evaluation methods for the web: A systematic mapping study. *Inf. Softw. Technol.* **53**, 789–817 (2011)
4. M. Fowler, *Refactoring: Improving the Design of Existing Code* (Addison-Wesley Longman Publishing Co., Inc., Boston, 1999)
5. A. Garrido, G. Rossi, D. Distanto, Refactoring for usability in web applications. *IEEE Softw.* **28**, 60–67 (2011)
6. J. Grigera, A. Garrido, J.M. Rivero, A tool for detecting bad usability smells in an automatic way, in *Web Engineering*, ed. by S. Casteleyn, G. Rossi, M. Winckler, (Springer International Publishing, Cham, 2014), pp. 490–493
7. J. Grigera, A. Garrido, J.M. Rivero, G. Rossi, Automatic detection of usability smells in web applications. *Int. J. Hum. Comput. Stud.* **97**, 129–148 (2017)
8. P. Harms, J. Grabowski, Usage-based automatic detection of usability smells, in *Human-Centered Software Engineering*, ed. by S. Sauer, C. Bogdan, P. Forbrig, R. Bernhaupt, M. Winckler, (Springer Berlin Heidelberg, Berlin, Heidelberg, 2014), pp. 217–234

9. P. Harms, S. Herbold, J. Grabowski, Trace-based task tree generation, in *Proceedings of the Seventh International Conference on Advances in Computer- Human Interactions (ACHI 2014)*, (CiteSeer, 2014)
10. F. Hermans, M. Pinzger, A.v. Deursen, Detecting and visualizing inter-worksheet smells in spreadsheets, in *Proceedings of the 34th International Conference on Software Engineering, ICSE '12*, (IEEE Press, 2012), pp. 441–451
11. L. Paganelli, F. Paternó, Tools for remote usability evaluation of web applications through browser logs and task models. *Behav. Res. Methods Instrum. Comput.* **35**, 369–378 (2003)
12. F. Paternó, A.G. Schiavone, P. Pitardi, Timelines for mobile web usability evaluation, in *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '16*, (Association for Computing Machinery, New York, 2016), pp. 88–91
13. F. Paterno, A.G. Schiavone, A. Conti, Customizable automatic detection of bad usability smells in mobile accessed web applications, in *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '17*, (New York, 2017)
14. R.F. Ribeiro, M. de Meneses Campanha, P.A.M.d.O. Souza, P. de Alcântara dos Santos Neto, Usability problems discovery based on the automatic detection of usability smells, in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC '19*, (New York, 2019), pp. 2328–2335
15. P. Weichbroth, Usability of mobile applications: A systematic literature study. *IEEE Access* **8**, 55563–55577 (2020)
16. C. Wohlin, P. Runeson, M. Hst, M.C. Ohlsson, B. Regnell, A. Wessln, *Experimentation in Software Engineering* (Springer Publishing Company, Incorporated, New York, 2012)

Part VII

High Performance Computing Architectures

Yun Tian, Saqer Alhloul, Fangyang Shen, and Yanqing Ji

Abstract

The student retention in undergraduate computer science degrees has been decreasing over the past 15 years where an attrition rate as high as 30% to 40% was observed during that period, with most students leaving the field after taking some introductory courses. Observing a similar trend or worse could be possible during or after the COVID-19 pandemic. A possible solution for this problem is to introduce highly motivating topics in the lower-division courses that will keep the students intrigued and wanting to learn more, and eventually, register in higher division courses. In this paper we provide a systematic review of the topics and subjects adopted in many institutions to motivate students of computer science in lower-division undergraduate curriculum. Then, we summarize the common attributes of these motivating subjects. In addition, we propose to leverage data processing subjects or big data problems to motivate college students to learn in lower-division courses. Based on the attributes of the existing motivating topics and the method of logic inference, we show that it is feasible and effective to motivate students' learning by injecting data processing subjects or big data problems in lower-division computer science courses.

Y. Tian (✉) · S. Alhloul
 Department of Computer Science and Electrical Engineering, Eastern Washington University, Spokane, WA, USA
 e-mail: ytian@ewu.edu; sahloul@ewu.edu

F. Shen
 Department of Computer System Technology, New York City College of Technology, Brooklyn, NY, USA
 e-mail: fshen@citytech.cuny.edu

Y. Ji
 Department of Electrical and Computer Engineering, Gonzaga University, Spokane, WA, USA
 e-mail: ji@gonzaga.edu

Keywords

Motivating students · Motivating topics · Big data · University retention · Lower-division curriculum · Computer education · STEM education · Motivational science · Data processing · Parallel computing

37.1 Background

Instructional and educational research showed that a decline in student enrollment in the field of computer science (CS) occurred in 2007, as well as between 2000 and 2005. An attrition rate as high as 30% to 40% had been observed during that time, with most students leaving the field after taking some introductory courses [1–3]. In 2013, a report showed that nearly half of college students who pursue science and technology degrees left the field or dropped out. And among all STEM fields, the attrition rate is highest for bachelor's degree candidates who declared a major in computer or information sciences and for associate degree candidates who declared a major in mathematics [4].

It is quite likely that this trend continues to grow in many institutions amid the ongoing COVID-19 pandemic, as the “global education emergency” threatens to derail the education of at least 24 million students projected to drop out of school as a result [5]. Therefore, it becomes ever more crucial and urgent to improve retention rates in CS, to provide an adequate number of skilled graduates to supply the needed workforce in today's economy. Motivating students to learn is one of the important approaches for improving students' retention and for providing high quality of education in academic institutions. Leutenegger and Edgington have pointed out, that attracting and retaining majors in computer science is all about engaging students with interesting assignments

and topics. If the assignments truly interest students then the probability of learning success and retention increases [2].

The latter conclusion motivated us to run a systematic review on the different subjects or topics used by many institutions to better engage lower division CS students. This is followed by extracting the common features from these topics which allowed us to provide a vision for how big data topics can be used as a motivating subject at lower division courses to increase retention rates and learning.

The rest of the paper is organized as follows. In Sect. 37.2, we review a group of existing motivating topics or subjects. Section 37.3 presents the common attributes of these motivating subjects. We argue that big data is effective and useful in motivating students' learning in Sect. 37.4. Finally, we conclude in Sect. 37.5 and comment on future research opportunities.

37.2 Existing Motivating Topics and Subjects

In this section several different motivating subjects that have been widely adopted are discussed. All these subjects have been effective to some extent to stimulate and retain students in introductory programming classes.

37.2.1 Robotics

One of the classical topics used for increasing retention is robotics topics [6–12].

Research suggested that to some extent students in robotics classes were motivated because they are interested in the content of the materials. Summet et al. argued that in their experiments, on average, the CS1-Robots students did 10% better than the CS1-NonRobots students, and success rate is higher for CS1-Robots students [11].

Klassner et al. proposed that the motivation to learn computing principles increases significantly when students have opportunities to apply those principles in constructing robots and designing problem-solving code [6]. At least seven core knowledge areas in CS can be motivated through robotics-related projects, including programming fundamentals, programming languages, algorithms and complexity, architecture, operating systems, intelligent systems, and net-centric computing [6]. However, the authors did not experimentally verify to what extent robotics topics may motivate students' learning in a classroom setting.

McWhorter et al. quantitatively and qualitatively verified how effectively robotics topics may motivate students' learning by using the Motivated Strategies for Learning Questionnaire (MSLQ) [10], where MSLQ can measure various aspects relating to student motivation, such as intrinsic

goal orientation, task value, control of learning beliefs, self-efficacy, and test anxiety [13]. In McWhorter's work, the quantitative data collected through MSLQ showed that the use of LEGO Mindstorms robotic activities had little if any impact on student motivation. Nevertheless, responses to the open ended follow-up questions suggested that the students really enjoyed the LEGO Mindstorms activities [10].

Kay et al. proposed to use robots in introductory courses as recruitment tools in computer science. The idea was based on the observation that the use of robots in introductory classes is an effective way to create an initial interest in CS and recruit students into their classes [12].

37.2.2 Game Programming

Game programming is another subject that is widely adopted in introductory programming classes to boost interests, because of its built-in motivation and familiarity for most students [2, 3, 14–17].

Leutenegger et al. believed that game programming motivates most new programmers. Interesting assignments mean that students are far more likely to learn because they are interested, and the visual component in game programming allows students to see mistakes in their code in a graphical way. This research showed that game programming improved student understanding of all seven basic topics examined [2].

Barnes et al. developed a game to teach introductory computer science concepts, called Game2Learn, to increase student motivation and engagement in learning CS1. The students' feedback demonstrated Game2Learn is effective in motivating students' learning because the students could have fun programming within a game, and in-game rewards and punishments are very important attributes that can motivate the students [3].

Mathew et al. used a Prosolve game to improve problem solving skills of beginning programmers in an introductory programming course. The feedback of the students demonstrated that the game helped a majority of the students understand the programming concepts and structures, as well as problem solving strategies. The authors also suggested that the game is a useful learning approach in attracting students' interest in the programming domain [17].

37.2.3 Simple 2D or 3D Graphics Libraries

Similar to game programming in which students are able to visualize the effects of their programming, simple 2D or 3D graphics libraries have been employed to teach fundamental programming concepts [18–22].

Roberts designed and implemented a simple 2-dimensional graphic library in C programming language for a CS1 course. The author's experience over the last two years at Stanford University suggested that using a graphics library in an introductory course enhanced student interest and helped reinforce several essential programming concepts [18].

Cooper et al. used 3D animated 3-dimensional graphics in a proposed CS1 course to introduce object-oriented programming concepts and help students develop problem-solving skills. The research results suggested the students in the class expressed a sense of self-confidence in their programming skills and the students demonstrated a high level of involvement [20].

Duarte et al. adopted Python with its Turtle Graphic Library in order to increase the interest and motivation of students in an introductory programming course. The research showed that students were engaged and motivated themselves with the graphical component [22]. Educators also taught Java programming with Graphics in CS1. It was pointed out that Object-Oriented concepts can be expressed most easily using graphics [19].

37.2.4 Other Motivating Topics

Hussain et al. used big data problems as motivational tools for initiating students in the area of high-performance parallel computing, as data sets become too large to fit the main memory of a single computer. For example, the authors adopted practical examples such as machine learning over large data sets in order to show the real-world impact of parallel algorithms [23]. The authors discussed the choice of the course materials and implementation strategies. However, no evaluation data was reported as to how effectively the curriculum materials motivated and engaged students.

Educators also incorporated parallel computing topics in lower-division computer science courses [24], although the benefits of such early adoption are still under investigation. We can envision that early adoption of parallel and distributed computing topics tends to motivate students to connect concepts in computer science with real-world problem solving, because most tasks in practice are parallel in nature.

37.3 Common Attributes of the Motivating Topics

In this work, we have no intention to compare or re-valuate these subjects to which extent they could motivate students. Whereas, by summarizing the above mentioned literature, we draw several quite enlightening conclusions concerning the common attributes found in these particular subjects. First,

these subjects are able to display the effects of programming in a visualized and visible form, so they tend to motivate students more aggressively. This visualizability of the outcome of students' program provides tangible and visible feedback of their coding, just as a photograph conveys information in a more vivid and more interesting fashion than lexical text.

Secondly, the above mentioned subjects promote the awareness of usefulness and importance of the contents. Robotics, game and graphics are closely relevant to industry or research projects. In particular, robotics topics may remind students of Mars' explorer and autonomous vehicles, while game and graphics are widely applied in movie production industry. Considering the ultimate goal of computer science education that prepares students professionally for a job, these subjects tend to be perceived to be useful and to be considered central to a student's identity, because of their real-world, applied and industry-related attributes. In other words, it is believed that acquiring these skills is crucial to students who ardently seek a career in the computer discipline.

According to findings in motivational science, although interest and intrinsic motivation can certainly motivate students to learn, it also matters whether students care about or think the task is important or valuable, with task value beliefs defined in terms of four components—intrinsic interest, utility, importance and cost [25]. Scientists in achievement motivation research believe that the usefulness or attainment value of a task that an individual perceived will directly relevant to boosting motivation and their engagement in the task [26–28]. The real-world, usefulness, applied and industry-related attributes of the three stimulating subjects above are exactly congruous with the theories discovered in the motivational science, proving the three subjects are effective to motivate students to learn in the discipline of computer science.

In the third place, all three subjects have already gained popularity and familiarity among teenagers, before they attend college. Nowadays, most youth have played with some kind of robot toys in or before high school, while video games have become one of the favorite activities of American children [29]. Cartoon or 3D movies that display fancy graphical images is another favorite entertainment among youth as well. Prior to college education, students have already attained some *superficial* acquaintance of these subjects, which is perceived as a source of fun or interests.

How is this familiarity and popularity of the subjects related to interests that motivate students in college? In addition to the fantasy pictures in games and graphics, curiosity and mystery plays a significant role. It is generally accepted that curiosity reflects a human tendency to make sense of the world and that we are curious about things that are unexpected or that we cannot explain. Curiosity is stimulated

by an information gap in our existing knowledge [14, 30, 31], while mystery is an external attribute of the subject itself. Mystery is enhanced by incongruity of information, complexity, novelty, surprise and violation of expectations [32], incompatibility between ideas and inability to predict the future, and information that is incomplete or inconsistent [30]. In a nut shell, mystery evokes curiosity in the individual and curiosity is one of the primary factors that drive learning [14].

Fourthly, the three subjects have another common attribute—mystery. Not only could a game's theme be mysterious, but game software could be technically mysterious also for a student. For example, with elementary knowledge of computing and game theory, students are intrigued by the questions, such as how the game software manages multiple players who are logged into the same server? how the characters or figures in the game are tracked on the game map? The same argument is true for robotics and graphics subject. Students that do not understand how a robot is able to autonomously go to the base and recharge itself are readily motivated and attempt to grasp the mechanism behind the scene. The complexity, novelty and surprising functionalities of these systems will definitely create mystery for students.

37.4 Topics of Big Data for Motivating Students

37.4.1 Emerging Needs for Big Data

As our world becomes ever more data driven, scientists and companies are confronted with a data deluge problem. It is ever more difficult and significant to extract meaningful knowledge or patterns from enormous amount of data. For examples, systems at Facebook, the social networking giant, scans roughly 105 *terabytes* of data each half hour. Operations such as filtering through messages that are associated with terrorism or unlawful activities have been crucial to security and welfare of our society. For another example, in order to optimize its supply-sale chain, the retailer Walmart might constantly scrutinize all its transaction data, to aggregate customers' preferences, review comments, promotions and price changes in order to draw a conclusion to facilitate decision making.

Educators in colleges have to deliver revolutionized topics to adapt to the changes in our society. Thus, modern data processing and big data skills become mandatory for computer majors nowadays in order to provide competent workforce for our societal economy. For example, in 2012 CNN Money foresees that the job as *data scientists* will grow 18.7% in the next ten years [33]. In another report, it is estimated that more than 90% of companies use cloud or big data platforms

to store or process data [34]. Interestingly, the urgency of motivating students in computer science has been coexisting in parallel with the big data revolution.

37.4.2 Big Data Topics Effective to Motivate Students

In this section, we justify that the topics of data science and data systems have the same set of attributes as the topics of game, robotics and graphics, thus we may adopt data science topics as an instrument to motivate students. In this section, we describe four attributes of the data processing or big data subject.

With the technologies of graphics and graphical user interface (GUI), data processing results can be presented and visualized in a very vivid and descriptive way. For an example, as a fundamental problem, requiring students to programmatically draw a circle or a triangle on an image could immediately grasps students' interests, and provoke thinking and attempts to relate computer programming with mathematic knowledge. Mining a dataset collected by the Centers for Disease Control and Prevention (CDC), students could visually discern how a type of virus is spread across different continents on the earth. For another example, GUI allows users to interact with a 3D map and display the areas of deforestation derived from satellite images.

Data processing or big data have been considered a crucial skill and perceived useful and important by students. Companies such as Microsoft, Amazon, Google, Walmart etc. have constructed their own data center processing their sales and other operational data. As a pioneer in cloud computing, for example, Amazon built its data center not only to serve its own needs, but also lease storage and computation power to the general public as cloud services. Not mentioning numerous companies attempt to optimize their business model by digging into their sales data, thus to reduce operational cost. Pioneered by these influential IT companies, data systems and processing have gained great momentum in today's society, and have permeated into the curriculum of some universities. Students have already recognized this trend as an external motivation, accepting that acquiring these skills helps land a decent job.

Data processing and big data systems are not unfamiliar among the youth. Research shows that students heavily use Facebook. More than 95% of students have heard about Facebook, 84% are members and at least 70% use the social network daily [35]. Data at FaceBook collected more than 500 terabytes per each day, with 2.5 billion pieces of content added each day [36]. Another example is Youtube. A survey in 2006 suggested that the YouTube website has 20 million visitors per month, with an additional 65,000 new videos uploaded every 24 h. More than 10 million video clips are

viewed daily on YouTube, with the most prominent age range being 12 to 17 years old [37]. Same as the subjects of robotics, game and graphics, college students have already been acquainted with these big data systems.

Likewise, the existing data systems are pieces of mystery for young college students. Students use these system regularly, nevertheless they rarely understand the underlying technology. For example, how does the Facebook data center store and process the gigantic volume of data? How could the website serve more than several million users at the same time without delay? From a different perspective, the data collected by a company also buries mystery in it. For instances, how does people's posting on FaceBook reflect their political opinion and affect a presidential election? Why the way of placing the commodities at Walmart is correlated with customers' purchasing patterns? The mystery in big data systems contributes to curiosity and stimulate learning as well.

With those stimulating attributes inherent in the data science subjects, in this work, we choose to leverage data processing systems as a tool to motivate students in low-division undergraduate courses. Similar to another pilot project, the University of California-Berkeley has launched program that engages youth (PreK-12 students) in hands-on, interactive data science projects, allowing youth to collect, map and visualize data [38]. That program attempts to attract more talented youth into the STEM fields. With a similar purpose in mind, in this work we try to teach data processing topics as a motivative means to attract more students into computer science, before they actually declare their major in college.

37.5 Conclusion and Future Work

This paper first provides a systematic review of the topics and subjects adopted to motivate students of computer science in Lower-division undergraduate curriculum. Next, we successfully summarize the common attributes of these motivating subjects. Then, in this work we propose to leverage data processing subjects or big data problems to motivate college students to learn in lower-division courses. Based on the attributes of the existing motivating topics in computer science, we are able to draw a conclusion that it is feasible and effective to motivate student learning by incorporating data processing subjects or big data problems in lower-division computer science courses. In the future, we will discuss choosing big data curriculum materials and how to deliver the topics in a lower-division course. Then we will verify its effectiveness through instructional surveys.

Acknowledgments This work is partially inspired and supported by the NSF IEEE TCPP Early Adopter grant to Eastern Washington University.

References

1. J. Vegso, Continued drop in CS bachelor's degree production and enrollments as the number of new majors stabilizes. *Comput. Res. News* **19**(2), 4 (2007)
2. S. Leutenegger, J. Edgington, A games first approach to teaching introductory programming, in *ACM SIGCSE Bulletin*, vol. 39(1) (ACM, New York, 2007), pp. 115–118
3. T. Barnes, E. Powell, A. Chaffin, H. Lipford, Game2learn: improving the motivation of CS1 students, in *Proceedings of the 3rd International Conference on Game Development in Computer Science Education* (ACM, New York, 2008), pp. 1–5
4. STEM-ming the Tide (2013) [Online]. Available: <https://www.insidehighered.com/news/2013/11/27/study-tracks-attrition-rates-stem-majors>
5. At least 24 million students could drop out of school due to the coronavirus pandemic, UN says (2020) [Online]. Available: <https://www.cnbc.com/2020/09/15/at-least-24-million-students-could-drop-out-of-school-due-to-the-coronavirus-un-says.html>
6. F. Klassner, S.D. Anderson, Lego mindstorms: not just for k-12 anymore. *IEEE Robot. Autom. Mag.* **10**(2), 12–18 (2003)
7. M. Guo, L. Husman, N. Vullum, A. Friesel, Project in robotics at the copenhagen university college of engineering, in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 2 (IEEE, New York, 2004), pp. 1375–1380
8. L. Greenwald, D. Artz, Y. Mehta, B. Shirmohammadi, Using educational robotics to motivate complete AI solutions. *AI Mag.* **27**(1), 83 (2006)
9. M. Petre, B. Price, Using robotics to motivate 'back door' learning. *Educ. Inf. Technol.* **9**(2), 147–158 (2004)
10. W.I. McWhorter, B.C. O'Connor, Do lego® mindstorms® motivate students in CS1? *ACM SIGCSE Bull.* **41**(1), 438–442 (2009)
11. J. Summet, D. Kumar, K. O'Hara, D. Walker, L. Ni, D. Blank, T. Balch, Personalizing CS1 with robots, in *ACM SIGCSE Bull.* **41**(1) (ACM, New York, 2009), pp. 433–437
12. J.S. Kay, Robots as recruitment tools in computer science: the new Frontier or simply bait and switch? in *AAAI Spring Symposium: Educational Robotics and Beyond* (2010)
13. C.R. Jackson, Validating and adapting the motivated strategies for learning questionnaire (MSLQ) for stem courses at an HBCU. *Aera Open* **4**(4) (2018). <https://doi.org/10.1177/2332858418809346>
14. R. Garris, R. Ahlers, J.E. Driskell, Games, motivation, and learning: a research and practice model. *Simul. Gaming* **33**(4), 441–467 (2002)
15. M. Kölling, P. Henriksen, Game programming in introductory courses with direct state manipulation. *ACM SIGCSE Bull.* **37**(3), 59–63 (2005)
16. R. Lawrence, Teaching data structures using competitive games. *IEEE Trans. Educ.* **47**(4), 459–466 (2004)
17. R. Mathew, S.I. Malik, R.M. Tawafak, Teaching problem solving skills using an educational game in a computer programming course. *Inf. Educ.* **18**(2), 359–373 (2019)
18. E.S. Roberts, A C-based graphics library for CS1. *ACM SIGCSE Bull.* **27**(1), 163–167 (1995)
19. S. Schaub, Teaching java with graphics in CS1. *ACM SIGCSE Bull.* **32**(2), 71–73 (2000)
20. S. Cooper, W. Dann, R. Pausch, Using animated 3d graphics to prepare novices for CS1. *Comput. Sci. Educ.* **13**(1), 3–30 (2003)
21. S. Matzko, T.A. Davis, Teaching CS1 with graphics and C. *ACM SIGCSE Bull.* **38**(3) (ACM, New York, 2006), pp. 168–172

22. E. Vidal Duarte, Teaching the first programming course with python's turtle graphic library, in *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education* (2016), pp. 244–245
23. F. Hussain, N. Deo, S.K. Jha, Early adoption: high-performance computing for big data introducing parallel programming and big data in the core algorithms curriculum, in *Proceedings of the 4th NSF/TCPP Workshop on Parallel and Distributed Computing Education (EduPar 2014)* (2014)
24. Center for Parallel and Distributed Computing Curriculum Development and Educational Resources PDC Curriculum Early Adopter Grant and Summer Training Program (2020) [Online]. Available: https://tcpp.cs.gsu.edu/curriculum/?q=NSF_Cybertraining
25. P.R. Pintrich, A motivational science perspective on the role of student motivation in learning and teaching contexts. *J. Educ. Psychol.* **95**(4), 667 (2003)
26. J.S. Eccles, A. Wigfield, In the mind of the actor: the structure of adolescents' achievement task values and expectancy-related beliefs (1995). <https://doi.org/10.1177/0146167295213003>
27. A. Wigfield, Expectancy-value theory of achievement motivation: a developmental perspective. *Educ. Psychol. Rev.* **6**(1), 49–78 (1994)
28. A. Wigfield, J.S. Eccles, The development of achievement task values: a theoretical analysis. *Develop. Rev.* **12**(3), 265–310 (1992)
29. D.A. Gentile, P.J. Lynch, J.R. Linder, D.A. Walsh, The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *J. Adolesc.* **27**(1), 5–22 (2004)
30. T.W. Malone, M.R. Lepper, Making learning fun: a taxonomy of intrinsic motivations for learning. *Aptit. Learn. Instruct.* **3**(1987), 223–253 (1987)
31. G. Loewenstein, The psychology of curiosity: a review and reinterpretation. *Psychol. Bull.* **116**(1), 75 (1994)
32. D.E. Berlyne, *Conflict, Arousal, and Curiosity* (McGraw-Hill Publishing Company Ltd., New York, 1960)
33. IT Data Scientist (2012) [Online]. Available: <http://money.cnn.com/gallery/pf/2012/11/01/best-new-jobs-in-america/3.html>
34. 5th Annual Trend in Cloud Computing (2014) [Online]. Available: <https://www.comptia.org/content/research/5th-annual-trends-in-cloud-computing>
35. C. Lampe, N. Ellison, C. Steinfield, A face(book) in the crowd: social searching vs. social browsing, in *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work. CSCW '06* New York, NY, (ACM, New York, 2006), pp. 167–170
36. J. Constine, How big is facebook's data? 2.5 billion pieces of content and 500+ terabytes ingested every day. Retrieved August, vol. 31 (2012), p. 2012
37. S.C. Burke, S. Snyder, R.C. Rager, An assessment of faculty usage of youtube as a teaching resource. *Internet J. Allied Health Sci. Pract.* **7**(1), 8 (2009)
38. Collaborative Proposal: From Data To Awesome (D2A): Youth Learning to be Data Scientists (2015) [Online]. Available: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1513296

Evaluation of Power Consumption and Application Optimization for Adaptive-Ticks Feature in Linux Kernel

Abdullah Aljuhni, Shaji Yusuf, C. Edward Chow, Oluwatobi Akanbi, and Amer Aljaedi

Abstract

Scheduler timer architecture has significant impact on operating system performance and power consumption. The current generation of Linux kernel supports multiple timer implementations, including periodic ticks, Dyntick-idle and Adaptive-ticks. Adaptive-ticks kernel offers the benefits of previous generations with additional improvement in power consumption and performance. In this paper, we evaluate the impact of Adaptive-ticks on power consumption with Linux kernel version 5.4.0 on an Intel Core i9-9900K. The current generation of Adaptive-ticks feature does not support multiple tasks in a ready queue; however, with the increase in application parallelism, not having support for multiple tasks in the ready queue poses a significant disadvantage to this feature. To support multi-threaded applications, we propose an application optimization technique which splits threads into two main categories, lightweight and heavyweight, with proper affinity settings for better power consumption. In addition, this study proposes a possible implementation strategy to extend the Adaptive-ticks feature to support multiple tasks in the ready queue. Our tests use in-band “RAPL” for profiling power consumption, and synthetic benchmarks such as Livermore, RAMSpeed, and SysBench, as the workloads. For real-world application benchmarking, we use Linux

kernel compilation. The study shows that Adaptive-ticks kernel can reduce power consumption by 1–2.7% and the application optimization technique provides a 2.4% enhancement in power consumption.

Keywords

Adaptive-ticks · Dyntick-idle · Full tickless · Scheduling timer · Power consumption · Performance

38.1 Introduction

Implementing high-resolution timer support in modern APICs allows for greater flexibility in control of the timer tick interval [1]. Operating systems use timer interrupts to manage multiple activities, including task and thread scheduling, software timers, and process preemption. The 2.6 series of Linux kernel, introduced in 2003, was the first to incorporate the support of a high-resolution timer into the mainline kernel (ver. 2.6.21) [2].

Linux kernel supports three different types of scheduling timers, including the periodic tick timer, Dyntick-idle timer, and Adaptive-ticks timer. Periodic tick timer was the legacy scheduler timer implementation, referred to in the kernel configuration setting as `CONFIG_HZ_PERIODIC` [3]. A periodic tick timer runs at fixed intervals and does not stop even during system idle time. This impacted the duration of sleep when the system was idling, as the timer ticks would wake up the system from sleep and also impact the ability of the processor to enter deeper sleep states. Dyntick-idle fixes the above flaw of periodic tick timer by switching off the timer when the system is in idle state so that it can enter sleep state, this is referred to in kernel configuration as `CONFIG_NO_HZ_IDLE` [3]. Dyntick-idle does not affect the performance of the system but improves the power

A. Aljuhni (✉) · C. E. Chow · O. Akanbi
Department of Computer Science, University of Colorado, Colorado Springs, CO, USA
e-mail: aaljuhni@uccs.edu; cchow@uccs.edu; oakanbi@uccs.edu

S. Yusuf
Intel Corporation, India Dev Center, Bangalore, India
e-mail: shaji.yusuf@intel.com

A. Aljaedi
College of Computing and Information Technology, University of Tabuk, Tabuk, Saudi Arabia
e-mail: aaljaedi@ut.edu.sa

consumption by enabling the processor to enter deeper and longer sleep states, when possible [4]. The most recent upgrade to scheduler timer mechanism, Adaptive-ticks, referred to in kernel configuration as `CONFIG_NO_HZ_FULL` [3], extends the ability of kernel to switch off the scheduling timer when there is only one task in the scheduler's ready queue.

The ability of Adaptive-ticks timer to extend the timer interval to a maximum time of 1 second when there is only one task in a ready queue ensures the system is not interrupted by timer when there is no need for scheduler to interrupt the task that is currently running. This new timer implementation is expected to improve both performance and power consumption by avoiding unwanted interrupts, which saves CPU cycles, memory bandwidth, and cache misses. CPU and memory bound workloads that execute for longer durations will take advantage of this type of timer. Real-time operating systems need the timer to tick at higher frequency to improve latency; due to this, Adaptive-ticks will have a greater degree of benefit for real-time systems.

Timer tick frequency can be set between 100 and 1000 Hz during the Linux kernel compilation. Desktop systems typically use a range of 300–1000 Hz, servers use a range of 100–300 Hz, and real-time systems run on 1000 Hz [5]. In server systems, long-timer interval will cause high latency, but server applications usually don't require low latency. On the other hand, short-timer interval is necessary for low latency in desktop systems and real-time systems.

The intensive power consumption of CPU workloads and CPU performance can be improved with Adaptive-ticks kernel by enabling non-preemptive scheduling and setting timer frequency at 100 Hz. This will reduce the number of interrupts, which will enhance execution time and cache efficiency [6]. However, this is not the best configuration for generic workloads because of the slow response to the events list [7]. Also, running multiple applications at the same time makes it hard on the operating system to predict tasks, which means the OS has no choice but to handle the events periodically as soon as they are in ready state [8].

Studying the impact of Adaptive-ticks on application performance and system behavior was part of the preliminary effort to understand the impact of this feature on performance [6]. However, the impact of Adaptive-ticks on power consumption was yet to be evaluated. Heavily threaded applications may find it difficult to make use of the current Adaptive-ticks feature because of the lack of multiple tasks support. Also, the current Adaptive-ticks feature lacks support for multiple tasks in the ready queue and hence scheduler timer design must be updated to support this feature.

The contributions of this paper are as follows:

- Studies the impact of Adaptive-ticks on system power consumption compared to the periodic tick kernel.
- Presents an application optimization technique for multi-threaded applications on Adaptive-ticks kernel for better power efficiency.
- Proposes a possible implementation strategy of Adaptive-ticks feature that includes support for multiple tasks in the task queue.

The remainder of this paper is organized as follows. Section 38.2 discusses the related work. Section 38.3 details the problem. Section 38.4 describes the system configurations. Section 38.5 explains the test methodology. The results of the experiment are presented in Sect. 38.6. Next, Sect. 38.7 details the implementation of the application's optimization technique. A proposed implementation strategy is shown in Sect. 38.8. Section 38.9 highlights our plans for future research. Finally, Sect. 38.10 presents the conclusion.

38.2 Related Work

Akkan et al. [9] presented a study focused on OS noise, especially in high-performance computing. They use a soft partition system to divided CPU's cores into OS cores and application cores. The operating system noise reduces with the Adaptive-ticks kernel. That should hold good only if we have one task in the ready queue. However, supporting multiple tasks is not addressed by their solution. Another relevant study was that of Siddha et al. [10], where the authors focused primarily on the effect of kernel latency by Adaptive-ticks operation. They found an improvement in kernel latency compared to the periodic tick operation. The efforts are to reduce the background activity on "tickless" operation during idle while our study focuses on Adaptive-ticks during load. Simonović et al. [11] focused on the power consumption of the Dyntick-idle system, specifically in terms of sleep state power consumption. In contrast to this study, our research focuses on power consumption during load. A study by Jiménez et al. [12] compared the number of interrupts that occur in the operating system with Dyntick-idle Linux kernel versus with the traditional periodic tick kernel. Their work achieved promising results. However, again, our focus is on Adaptive-ticks kernel during load. Furthermore, Garcia et al. [13] presented various optimization techniques to improve power consumption on Power6 system in idle mode. In the

optimization, they used applications, hardware, and operating system techniques. Adaptive-ticks kernel was used during idle as part of the optimization; but power consumption during active mode was not analyzed by the Garcia et al. study.

38.3 Problem and Motivation

The study of power consumption of various kernel features provides significant data points to understand how to achieve maximum performance with the lowest possible power consumption. Adaptive-ticks kernel is known to provide better performance in most CPU and memory bound workloads, however the impact of this on power consumption is not completely analyzed. This study tries to fill this gap by focusing primarily on understanding the CPU power consumption under Adaptive-ticks kernel.

During the study of performance and power consumption, it became apparent that most modern applications implement parallelization to a large extent, which limits the current effectiveness of the Adaptive-ticks feature. However, in the current Adaptive-ticks implementation, the scheduler will check whether there is only one task in the ready queue and then reprogram the HR timer to run without interrupt for 1 second before hitting the next timer tick. Forcing the system to set thread affinity in such a way that all the heavyweight threads get scheduled in core 1 through core n, and having all the lightweight threads forced to be scheduled on core 0, would necessarily mean that the system runs identical to an Adaptive ticks kernel with only one task in the ready queue.

Application optimization for Adaptive-ticks kernel poses significant challenges in terms of redesigning applications that already exist. Linux kernel should support more than one task in the ready queue while running in Adaptive-ticks mode to get the best performance and power consumption for demanding modern workloads.

38.4 System Configurations and Settings

Adaptive-ticks was evaluated on a machine with Intel Core i9-9900K; base frequency 3.60 GHz with 16-MB L3 cache, motherboard model Z390 DESIGNARE-CF, Gigabyte Ltd.; 32-GB RAM Dual Channel, and Samsung SSD 840 Pro Disk. The operating system under test is CentOS 7, and the tested kernel version is 5.4.0. Table 38.1 lists the system configurations and boot parameters.

Table 38.1 System configurations

Dynsticks-idle Kernel – Config Param	Value
Dynticks-idle	ON
Timer frequency	1000 Hz
Adaptive-ticks Kernel – Config Param	Value
CPU timer accounting	Full dynticks accounting
Full dyntick system (Tickless)	ON
Full dyntick system on all CPUs	ON – (except CPU0)
Hight-resolution timer support	ON
Timer frequency	1000 Hz
Adaptive-ticks Kernel Boot Parameters	
nohz_full = 1–7, isolepus = 1–7, rcu_nocbs = 1–7, rcu_nocb_poll, irqaffinity = 0	

38.5 Experimental Design and Methodology

The experiments were designed to evaluate the differences in power consumption of two different kernel configurations. The first kernel was compiled with Adaptive-ticks feature. The second kernel was compiled with Dyntick-idle feature. We performed a Linux kernel compilation operation as real-world workload beside the selected micro-benchmarks to obtain comprehensive and accurate results.

To avoid unexpected performance behavior, Intel Turbo Boost Technology and Deep Sleep States are disabled [14]. In Linux, many governors have different policies on the control of frequency, voltage, and power states. To avoid mixed results and get accurate results, we choose CPUfreq userspace governor and the default governor Intel P-State.

The experiments focused on measuring the power consumption improvement during load based on two factors. The first factor was the improvement that takes place in cache performance and processor pipelines due to how replacing some cache lines with ISR can be avoided [15]. The second factor was focused on the saved CPU cycles that can be avoided as they serve no purpose.

To obtain accurate results, we performed a kernel compilation as CPU-intensive real-world workload. Also, as part of calculating the power consumption, we looked equally at the amount of consumed power, performance results, and execution time. Furthermore, since the expected power consumption improvement was between 1% and 3%, p-value for each experiment was calculated to show the significance of each test result. In our statistical analysis, the signifi-

cance level α was 0.05. Moreover, because there is no known standard deviation for each experiment, all tests were repeated at least 30 times. Both kernels, the Dyntick-idle kernel and the Adaptive-ticks kernel, were run on the same machine with little modification in the grub boot loader for boot parameters.

In the experiments, we mainly focused on two power management governor settings to evaluate the power consumption results. The first governor was Intel P-State, which is the default driver in Linux kernel. It is an advanced pair driver that controls both CPU voltage and frequency [16]. The second governor was CPUfreq Userspace. Userspace is supported by ACPI-CPUfreq and allows the CPU frequency to be modified dynamically from userspace. The CPUfreq Userspace governor clock speed was set to 3.60 GHz during our tests.

In our experiments, the focus was on calculating the total energy consumed by CPU cores. For that purpose, perf tool was used due to its accurate power reading and ease of merge with the tests. Perf is a powerful lightweight tool used widely in Linux kernel to measure various activities. Perf tool can be used to monitor and profile the power consumption at a level of granularity [17]. It depends on RAPL interface to log the power consumption results with a high level of accuracy from some MSRs (Model Specific Registers) on $\times 86$ systems [18].

In an automated way, a test script was used to run and log the results of the experiments. Other system parameters were monitored before and during the load, such as thermal throttling count, CPU load, CPU frequency, cache misses, and CPU temperature. All possible measures were taken to ensure the test environments for each experiment were as similar as possible.

38.6 Test, Result, and Analysis

In this section, the experimental results of four different types of workloads are presented and analyzed. The workloads were chosen carefully, based on our understanding of Adaptive-ticks implementation level. The observed system parameter indicates a stable test environment with “zero” CPU throttling, steady clock speed, and the expected number of interrupts with Adaptive-ticks kernel and Dyntick-idle kernel was in the normal standard.

38.6.1 Power Consumption of Livermore Loops

Livermore loops is a well-known benchmark tool used in supercomputers to assess performance. It consists of 24 kernels and performs sequence operations of floating-point/vector

[19]. The results are represented as mega floating-point operations per second (MFLOPS). However, our interest was in measuring the consumed power to perform the test.

Figure 38.1a shows the average power consumption obtained from running Livermore benchmark on Adaptive-ticks kernel versus Dyntick-idle kernel. In both kernels, all cores were running with Userspace governor, and the frequency speed was set to 3.60 GHz. The results show us a 1.4% reduction in power consumption with Adaptive-ticks kernel compared with that achieved by Dyntick-idle kernel. The experiment’s calculated p-value was $5.64E-59 < \alpha$. This indicates a statistical significance in the percentage of improvement in energy consumption. The enhancement in power consumption is due to the fact that some unnecessary CPU cycles are avoided by Adaptive-ticks kernel, and the cache is more optimized with this feature [6]. The performance for both kernels was identical, and the most gain was in reduction in power consumption.

Figure 38.1b illustrates the average power consumption obtained from running Livermore benchmark on Adaptive-ticks kernel against Dyntick-idle kernel with Intel P-State governor. There is a 1% enhancement in power consumption in favor of Adaptive-ticks kernel. The experiment’s p-value is $2.25E-43 < \alpha$, which shows statistical significance to the enhancement in power consumption. The average performance for both configurations were identical.

38.6.2 Power Consumption of RAMSpeed

RAMspeed is a benchmark tool used to measure cache efficiency and memory bandwidth for a single application with a large amount of memory. It generates a memory intensive workload with a lot of reads and writes operations to stress the cache and memory [20]. Among other testes available with RAMspeed, we selected the INTmem test due to its simplicity and similarity to the real-world application.

Figure 38.2a illustrates the average enhancement in power consumption of Adaptive-ticks kernel with Userspace governor compared to that of Dyntick-idle kernel with the same governor. In both kernels, CPU speed was manually set to 3.60 GHz. The results show improvement in power consumption of about 2.5% with Adaptive-ticks kernel. P-value for this experiment was $3.79E-33 < \alpha$, which indicates statistical significance in power consumption enhancement. Also, in terms of the average execution time, there was an enhancement of 1% with Adaptive-ticks kernel. Moreover, p-value for the average execution time was $5.40E-13 < \alpha$, which also indicates significant performance enhancement with Adaptive-ticks kernel compared to Dyntick-idle kernel.

Figure 38.2b shows an enhancement in average power consumption of about 1.6% in favor of Adaptive-ticks kernel

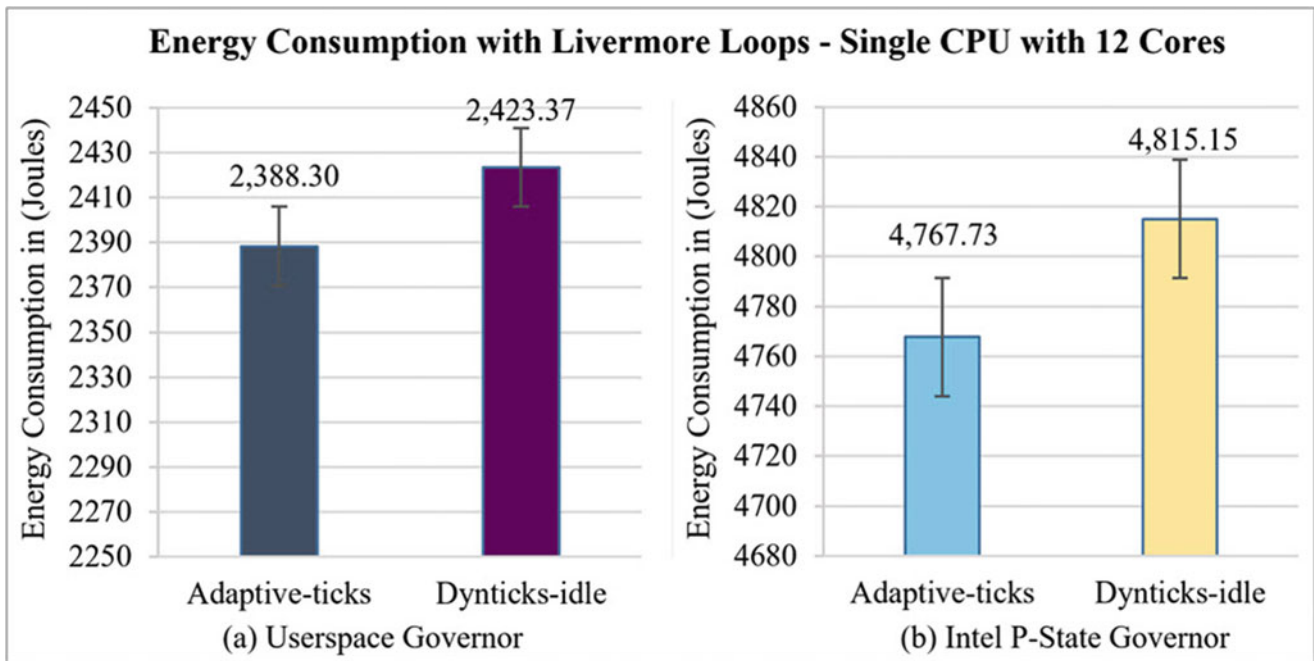


Fig. 38.1 Livermore – power consumption of Adaptive-ticks and Dyntick-idle timers. (a) Userspace governor. (b) Intel P-state governor

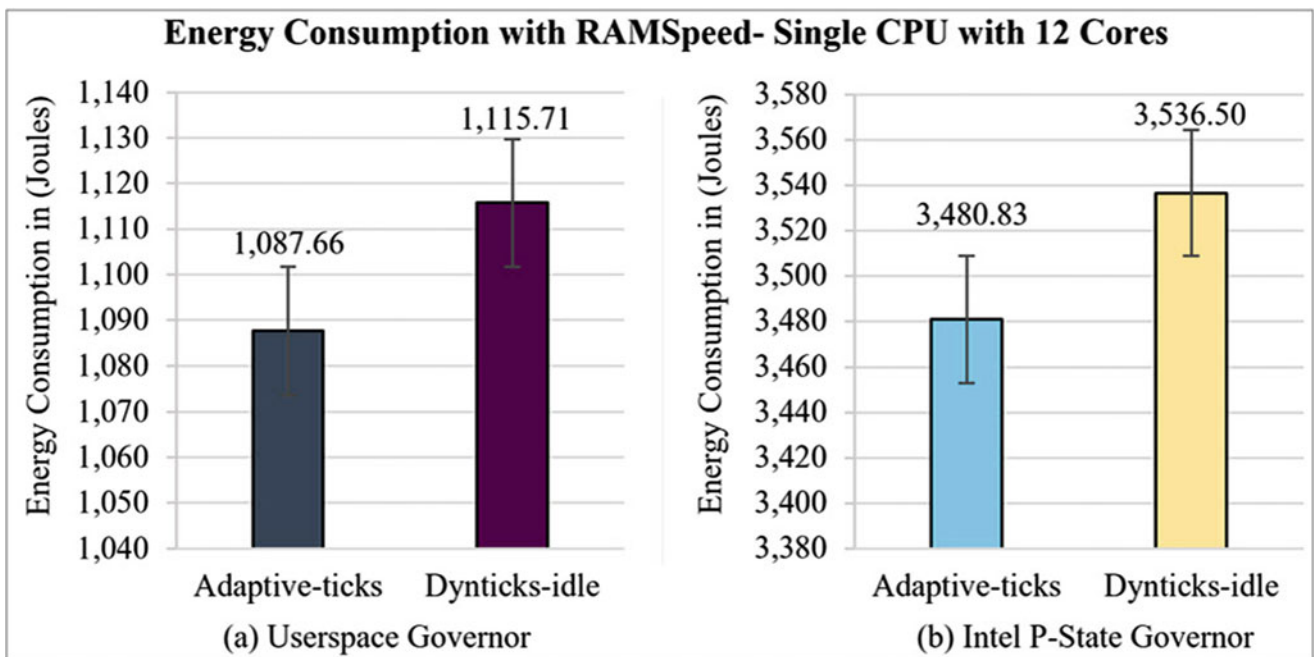


Fig. 38.2 RAMSpeed – power consumption of Adaptive-ticks and Dyntick-idle timers. (a) Userspace governor. (b) Intel P-state governor

with Intel P-State governor. P-value here was $0.012 < \alpha$, which indicates a statistically significant enhancement in power consumption. Also, on the other side of the test, there was an enhancement in the average execution time of 1.7%. In the average execution time, p-value was $1.145E-05 < \alpha$, which indicates a significant enhancement in performance.

38.6.3 Power Consumption of Linux Kernel Compile (GCC)

For real-world applications, we used the GNU C Compiler GCC version 7.5 to compile Linux kernel version 5.4.0. Compiling Linux kernel requires performance and complex

memory access as it is a large-size C language software project. A lot of operations are active during this compilation, such as compile, link, create symbol files, and generate a binary file. Real world applications give a good indication of performance and power consumption versus what is evident using a synthetic benchmark, which may sometimes give bias results. The compilation is done with default flag without parallel compilation.

Figure 38.3a illustrates the average enhancement in power consumption with Adaptive-ticks kernel using Userspace governor at 3.60 GHz, compared to that of Dyntick-idle kernel with the same governor and frequency speed. There is a 1.2% enhancement with Adaptive-ticks kernel. P-value for this experiment was $0.0043 < \alpha$, which indicates a statistical significance in the enhancement. Moreover, the average compilation time with Adaptive-ticks kernel was 2% better. P-value was $3.3E-05 < \alpha$ for the average execution time, which shows us significant enhancement in performance.

Figure 38.3b presents a similar experiment to that of Fig. 38.3a, only with Intel P-State power management governor. The results show enhancement in power consumption that favors Adaptive-ticks kernel by 1.1%. The experiment's p-value was $0.00045 < \alpha$, which indicates a statistically significant enhancement. In terms of performance, Adaptive-ticks kernel shows improvement of 1.2%, where the experiment's p-value was $1.7E-05 < \alpha$, which also indicates statistical significance in performance enhancement.

38.6.4 Power Consumption of Sysbench and MySQL

SysBench is widely used to observe and understand an operating system's performance and behavior under load. It is highly customizable to test different types of databases [21]. In our experiment, we used a built-in scripts called `oltp_read_write` to generate a workload with MySQL version 5.7.

Figure 38.4a shows the average power consumption for read and write transactions in MySQL 5.7. Adaptive-ticks kernel with Userspace governor and 3.60 GHz obtains 2.2% better power consumption compared to Dyntick-idle kernel with the same parameters. P-value for the experiment was $2.0E-06 < \alpha$, which indicates a statistically significant enhancement. On the other side, the number of transactions and the execution time for both kernels were identical.

Figure 38.4b illustrates enhancement in power consumption of about 2.4% toward Adaptive-ticks kernel with Intel P-State governor. P-value indicates a statistical significance in the experiment with $1.93E-05 < \alpha$. Also, similar to Fig. 38.4a, the results with the Intel P-State governor reveal no improvement in performance or execution time.

38.7 Application Optimization for Adaptive-Ticks

Heavily threaded applications may find it difficult to make use of the current Adaptive-ticks feature because of the lack

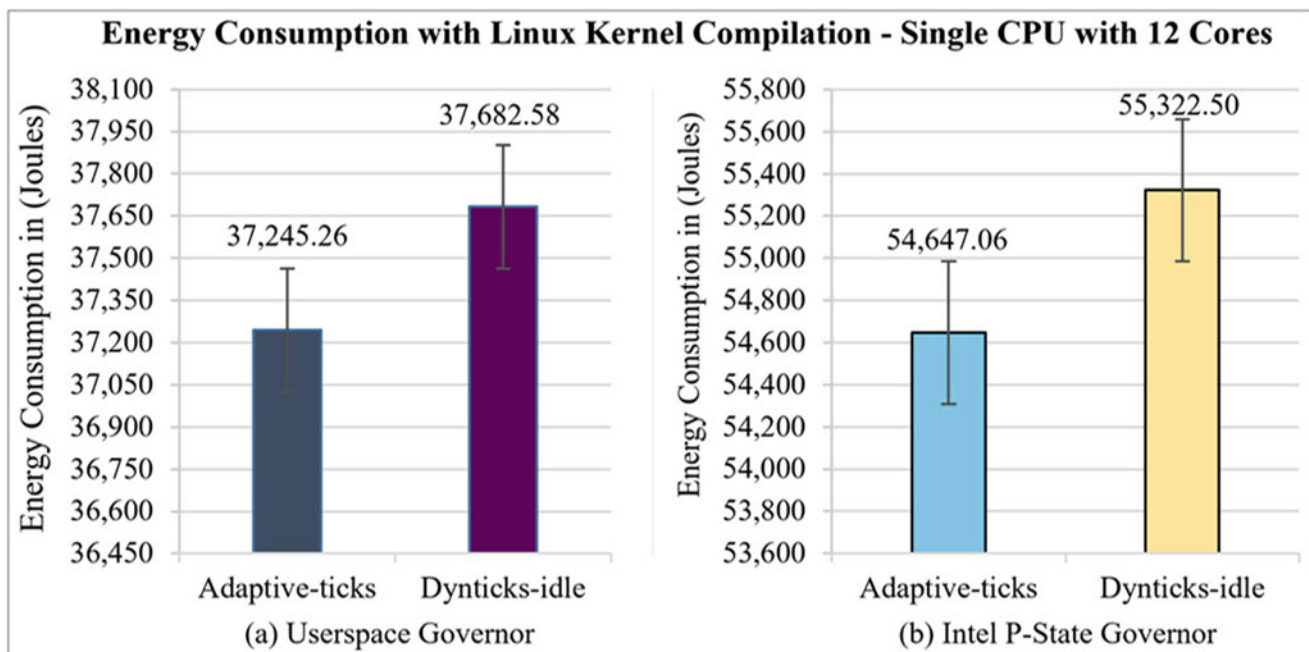


Fig. 38.3 Kernel compilation – power consumption of Adaptive-ticks and Dyntick-idle timers. (a) Userspace governor. (b) Intel P-state governor

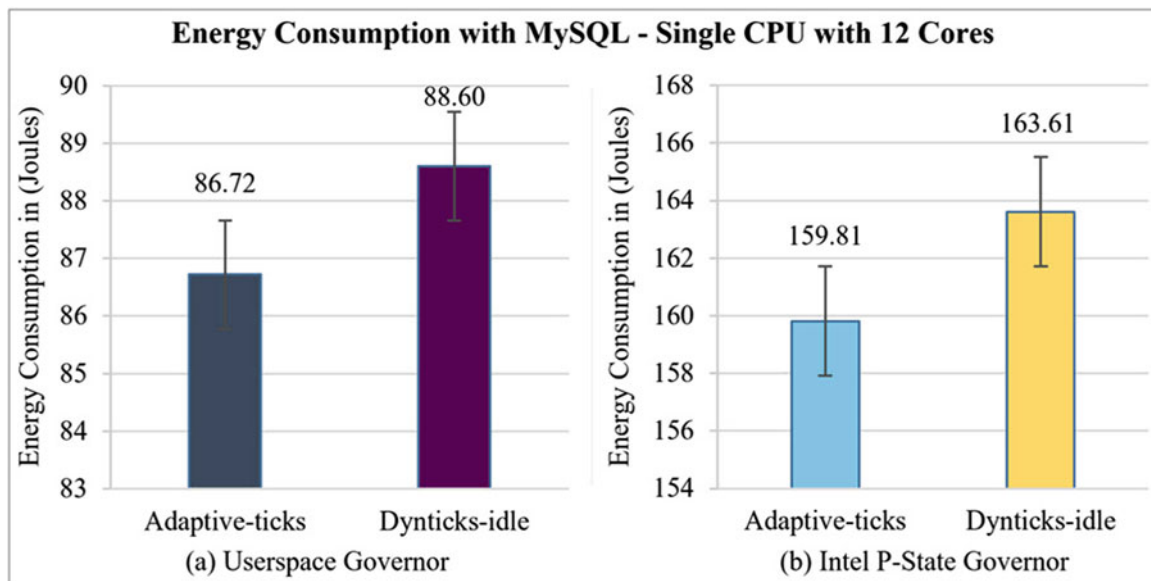


Fig. 38.4 MySQL – power consumption of Adaptive-ticks and Dyntick-idle timers. (a) Userspace governor. (b) Intel P-state governor

of multiple tasks support. Such applications can still benefit from the feature by redesigning the application in such a way that it is divided into highly CPU intensive threads and lightweight threads. All the lightweight threads should be set to schedule on core 0, and each of the rest of the threads should be allocated to the remaining cores. In this way, each of cores 1 through n will have only one thread allocated, whereas core 0 will have multiple light threads allocated to it. Having core 0 in periodic ticks mode will not impact performance when multiple threads are allocated to it.

The above can be implemented by designing the application in such a way that there are two types of threads in the application, one that is heavily CPU/memory intensive and the other which is lightweight threads that handle IO or other management tasks. We then set the affinity of the lightweight threads to core 0 and the heavyweight threads to Adaptive-ticks cores 1 through n using the `pthread_attr_setaffinity_np` function that maps threads to specific physical cores.

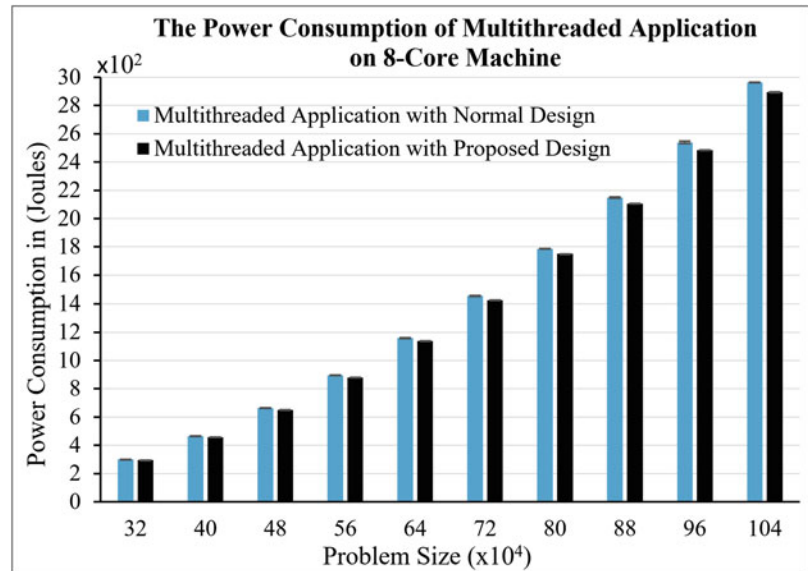
To verify our proposed solution, we built two multiple threaded applications to calculate the prime numbers and then set the affinity for threads, as explained above, in one program, while the other program has no CPU affinity set [22]. Inter thread communication and synchronization are set to the minimum to ensure the threads don't wait for synchronization. Threads are divided based on the type of workload, meaning highly CPU/memory intensive threads and lightweight threads that manage background activities. All the lightweight threads are scheduled on core 0, and each of the heavyweight thread is allocated on each of the other cores. Now, the heavy threads allocated on cores 1–7 can operate in Adaptive-tick mode because there is

only one thread assigned to each of those cores. However, the threads allocated to core 0 have minimal impact because of the periodic tick operation of core 0, as those threads only manage IO and other lightweight housekeeping activities.

Figure 38.5 presents a comparison of the power consumption between two multi-threaded applications, calculating prime numbers on the same Adaptive-ticks kernel. The first application is built based on the proposed design and the second application is built without optimization. The comparison shows enhancement in the power consumption of between 2% and 2.65% in favor of the application that employs the proposed design. Changing the problem size had no impact on the measured power consumption and the percentage enhancement is in the range of 2–2.65%. The p-value was calculated for each problem size and found to be in the range of $3.1E-10$ and $9.4E-13$, which indicates significant enhancement in power consumption. On the other side, the performance is found to be identical for the two applications.

Application developers can lower system power consumption by 2–2.65%, if the above findings are taken into consideration when building multithreaded applications for Adaptive-ticks kernel. Improvement in performance and power consumption for multi-threaded applications also depends on the nature of the problem, how far it can be parallelized, and whether its tasks can be split over a multi-core system with minimum interference. Because it gives a 2–2.65% enhancement in power consumption without negatively impacting performance, the overall performance per watt increases by an average factor of 2.4%.

Fig. 38.5 Multit-headed application – power consumption of Adaptive-ticks



38.8 Proposed Possible Implementation for Multiple Tasks Support on Adaptive-Ticks Kernel

The excessive power consumption of Linux, primarily attributed to the inability to enter and remain in deeper sleep states during idling in the desktop and mobile configurations, resulted in ‘Dyntick-idle’ implementation in kernel 2.6.21. Dyntick-idle feature of kernel takes the system to tickless mode (mode with very few ticks) during system idle time, which helps the system utilize CPU power management features more efficiently. Dyntick-idle feature is found to be very efficient in keeping system power consumption lower while idling with almost no performance benefit or degradation.

Turning off the timer is done during scheduling timer interrupt at the exit function. The system checks for the need to turn on timer at multiple locations, including timer interrupt exit function, context switching, IPI and IRQ work function.

Figure 38.6 explains the design of Dyntick-idle kernel that we are benchmarking against. Adaptive-ticks kernel is implemented by adding a condition check to the timer exit function of the Dyntick-idle kernel. To check the number of tasks in the ready queue and enable 1 second timer duration if only one task is available in the task queue.

Multiple task support is implemented primarily by including an extra function to determine the duration of the timer deferment interval for each task requires in the scheduler core of the kernel. For instance, if there is only one task in the ready queue, then just as in the current Adaptive-ticks kernel, the task will get 1 second of the time slice and there will be only one timer interrupt after that second and if there are two tasks in the ready queue, then the time slice will be divided between the tasks as 500 ms and there will be two timer interrupts per second, and so on. This is a simple

algorithm approach that does not consider the task priorities and the maximum number of the supported tasks in the ready queue. However, a detailed analysis of the new design and its interaction with the other parts of the kernel would result in better and more efficient algorithms to determine the right time slice interval in each timer interrupt for better performance and more efficient power consumption.

38.9 Future Work

Possible area for future study would include removing the timer from all but one or two cores is the primary goal. That would result in both better performance and power efficiency. The idea here is not to run timer in all the cores and to instead run the timer in only one core use IPIs to signal the other cores to start or preempt tasks in other cores. This would mean cores will be interrupted based on the decision of periodic tick core. Also, another enhancement for Adaptive-ticks kernel can be done in Xeon servers with support for Sub-NUMA Clustering [23]. This would involve applying memory affinity to the interrupt service routine (ISR) to the memory located in the right memory controllers, making sure that with each tick, the interrupt service routine is loaded from the closest memory controller. That would improve the latency and the response for real-time application, as well as reduce OS jitter. In addition, heterogenous CPUs can have significant enhancement with Adaptive-ticks feature due to the availability of big and small core clusters. The goal is to use one of the low-performance cores as a periodic tick core all the time instead of switching the periodic core on the high-performance cluster. This change should result in an increase in the performance by freeing up the high-performance core and at the same time, result in saving power by using the low performance core.

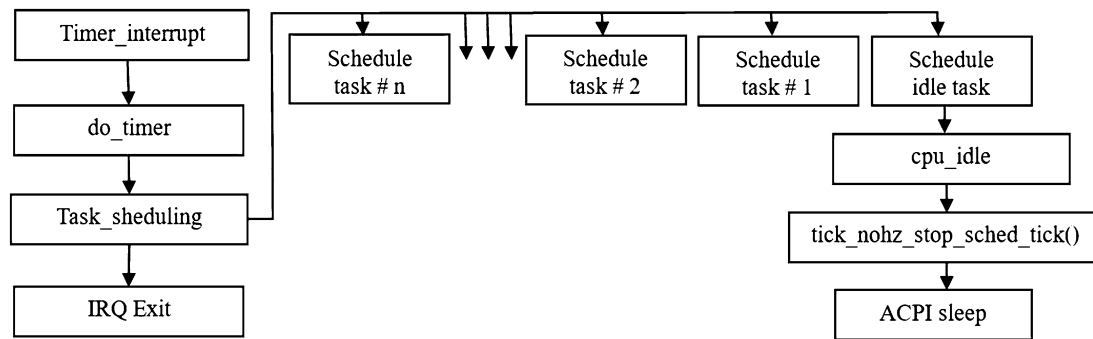


Fig. 38.6 Dyntick-idle – high level architecture

38.10 Conclusion

This study successfully illustrated the scale and scope of the effect of Adaptive-ticks kernel on system power consumption in contrast to that of the Dyntick-idle with the help of multiple benchmarks and workloads. The benchmarks selected represents realistic workload conditions to evaluate the power consumption on Adaptive-ticks kernel.

The Adaptive-tick timer shows positive impact on power consumption, though support for multiple tasks limits its scope to situations where only one task is available in the ready queue. Successful illustration of application optimization technique to extend the scope of this feature to applications that are highly thread optimized proved that the scope of this feature is extensible to heavily threaded applications as well.

The proposed implementation of multi task support extends the feature to cover the entire spectrum of applications whether heavily thread optimized or unoptimized. This implementation will make the feature compete in all aspects by covering all possible scenarios and does it without the need for redesigning applications.

Acknowledgments The main author would like to thank the expert in intel Mr. Shaji Yusuf for his insightful advices and discussion. This research is partially supported by Saudi Arabian Cultural Mission (SACM) in USA and Tabuk University.

References

1. Intel 64 Architecture X2apic specification. (2010, January 01). Retrieved April 22, 2020, from <https://software.intel.com/content/www/us/en/develop/download/intel-64-architecture-x2apic-specification.html>
2. T. Gleixner, D. Niehaus, Hrtimers and beyond: Transforming the Linux time subsystems, in *Proceedings of the Linux Symposium*, vol. 1, (2006)
3. NO_HZ: Reducing Scheduling-Clock Ticks, https://www.kernel.org/doc/Documentation/timers/NO_HZ.txt
4. J. Corbet, (Nearly) full tickless operation in 3.10, <https://lwn.net/Articles/549580/>, (May 2013)
5. F. Weisbecker, Status of Linux dynticks, in *9th Annual Workshop on Operating Systems Platforms for Embedded Real-Time Applications-OSPERT13*, (2013)
6. A. Aljuhni, C.E. Chow, A. Aljaedi, S. Yusuf, F. Torres-Reyes, Towards understanding application performance and system behavior with the full dynticks feature, in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, (IEEE, 2018), pp. 394–401
7. M. Holappa, Performance Comparison of LTE ENODEB OSI Layer 2 Implementations: Preemptive Partitioned Scheduling vs. Non-Preemptive Global Scheduling. Master's Thesis, Degree Programme in Information Networks (2013)
8. M. Palmer, M. Walters, *Guide to Operating Systems* (Cengage Learning, Boston, 2012)
9. H. Akkan, M. Lang, L.M. Liebrock, Stepping towards noiseless Linux environment, in *Proceedings of the 2nd International Workshop on Runtime and Operating Systems for Supercomputers*, (June 2012), pp. 1–7
10. S. Siddha, V. Pallipadi, A. Ven, Getting maximum mileage out of tickless, in *Proceedings of the Linux Symposium*, (Ottawa, 2007), pp. 201–207
11. M. Simonović, L. Saranovac, Power management implementation in FreeRTOS on LM3S3748. *Serbian J. Electr. Eng.* **10**(1), 199–208 (2013)
12. V. Jiménez et al., Power and thermal characterization of POWER6 system, in *2010 19th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, (IEEE, 2010), pp. 7–18
13. F. Garcia, C. Lameter, Tickless kernel practical experiences, in *Presented at LinuxCon*, (2013)
14. S. Kanev, K. Hazelwood, G.-Y. Wei, D. Brooks, Tradeoffs between power management and tail latency in warehouse-scale applications, in *2014 IEEE International Symposium on Workload Characterization (IISWC)*, (IEEE, 2014), pp. 31–40
15. D. Thiebaut, H.S. Stone, Footprints in the cache. *ACM Trans. Comput. Syst. (TOCS)* **5**(4), 305–329 (1987)
16. Intel pstate CPU performance scaling driver, https://www.kernel.org/doc/html/v4.12/admin-guide/pm/intel_pstate.html (n. d.)
17. A.C. de Melo, The new Linux 'perf' tools. Slides from Linux Kongress (2010)
18. E. Rotem, A. Naveh, A. Ananthakrishnan, E. Weissmann, D. Rajwan, Power-management architecture of the intel microarchitecture code-named Sandy Bridge. *IEEE Micro* **32**(2), 20–27 (2012)

19. F.H. McMahon, *The Livermore Fortran Kernels: A Computer Test of the Numerical Performance Range* (Lawrence Livermore National Lab, Livermore, 1986)
20. R.M. Hollander, P.V. Bolotoff, RAMSpeed, a cache and memory benchmarking tool (2011)
21. A. Kopytov, Sysbench manual. MySQL AB, 2–3 (2012)
22. A. Aljuhni, Optimized multithreaded, <https://github.com/UCCSCO/optimized-multithreaded> (2020)
23. A. Sodani, Knights landing (knl): 2nd generation Intel® Xeon Phi processor, in *2015 IEEE Hot Chips 27 Symposium (HCS)*, (IEEE, 2015), pp. 1–24

Sequence Alignment Algorithms in Hardware Implementation: A Systematic Mapping of the Literature

39

Lucas S. M. Bragança, Adler D. Souza, Rodrigo A. S. Braga, Marco Aurélio M. Suriani, and Rodrigo M. C. Dias

Abstract

One of the primary activities of the bioinformatics is the alignment of sequences, which can find similar patterns between two sequences and determine their structure, their control information, or even their own functions. The growth of databases increases the computational effort spent to execute sequence alignment algorithms. Consequently, it can take a considerably long time to run these algorithms on general purpose processors. This paper aims to map and analyze articles related to the implementation of sequence alignment algorithms in FPGAs and GPUs to identify the most recent findings on the subject, as well as possible gaps that may lead to further investigations. The systematic mapping led to the selection of twenty-three articles using FPGA and the GPU as the hardware platform. It also identified six sequence alignment algorithms: Needleman-Wunsch, Smith-Waterman, HMM, BLAST, BWA e FMIndex. The present work was able to evaluate how often these hardware and algorithms are being used in scientific researches and their benefits in terms of processing time and energy consumption.

Keywords

Bioinformatics · Sequence alignment · Algorithms · Hardware · Parallel processing · Smith-Waterman Algorithm · FPGA · GPU · Execution time · Review

L. S. M. Bragança · R. A. S. Braga (✉) · M. A. M. Suriani
R. M. C. Dias
Institute of Science and Technology, Federal University of Itajubá,
Itajubá, MG, Brazil
e-mail: lucasbraganca@unifei.edu.br; rodrigobraga@unifei.edu.br;
marcosuriani@unifei.edu.br; rmcdias@unifei.edu.br

A. D. Souza
Institute of Mathematics and Computing, Federal University of Itajubá,
Itajubá, MG, Brazil
e-mail: adlerdiniz@unifei.edu.br

39.1 Introduction

Bioinformatics is an interdisciplinary field interconnecting molecular biology and computing. Its main goal is to use computers to analyze and organize biological data in areas such as genome mapping and sequencing. Some of the main tasks of this field are to predict three-dimensional structures of either sequences, structures or biological/biophysical functions of sequences, as well as to simulate metabolism or other basic biological processes of these functions. Thus, this field provides a better comprehension of how certain diseases originate and how to create new drugs to combat them. Its applications also include phylogeny, metagenomics, and agriculture [1, 2].

According to [2], computational biology focuses on three key areas: (i) sequence alignment, (ii) structural analysis and (iii) function prediction. The structural analysis investigates biological structures to determine their sequence, function, and their control information. Function prediction seeks to understand how sequences and structures take specific functions. Finally, sequence alignment analyses DNA and protein sequences and tries to determine their structure, function, and control information.

Many algorithms have been developed to conduct these studies. Some of them can take advantage of parallel hardware processing, especially those with high complexity. Among the algorithms used to perform the sequence alignment are the Needleman-Wunsch [3] and the Smith-Waterman [4]. Both of them deal with large size data sets that demand a high expenditure execution time. However, for extensive data sets, they may not be able to be run on a CPU. Besides the execution time, some issues related to energy consumption must also be addressed when choosing an algorithm.

Thus, to discover the algorithms for sequence alignment implemented in FPGAs and GPUs, we propose a systematic

mapping of literature [5]. For the systematic mapping, we searched for articles about that topic published from 2015 to 2020 in three different scientific databases.

This manuscript is organized as follows, Sect. 39.2 describes the methodology and presents the research definition and questions, as well as the execution and data extraction from the selected articles. Section 39.3 presents the results and analyzes the data obtained from the articles. Section 39.4 discusses the results and answers the research questions. Finally, Sect. 39.5 concludes this manuscript.

39.2 Research Definition

A systematic mapping of the literature is a publication specialized in locating and selecting other scientific publications, and in analyzing their contributions. This sort of work establishes and utilizes a set of clear steps and criteria, allowing the authors to draw conclusions about the current state-of-art of a particular subject [5]. In the present systematic literature mapping, we carried out the following steps: (i) define the research questions, (ii) develop the research protocol, (iii) select the academic databases, (iv) apply the inclusion and exclusion criteria, (v) identify/analyze the selected studies and (vi) answer the research questions [5].

39.2.1 Research Questions

The purpose of this work is to map recent publications referring to sequence alignment algorithms implemented in FPGAs and GPUs. Thus, we define two research questions:

RQ1: Which sequence alignment algorithms have been implemented in FPGAs and GPUs in recent years? The key research question seeks to map the sequence alignment algorithms that are being implemented in FPGAs and GPUs platforms. In this way, it will be possible to investigate the possibility of implementing computational biology algorithms.

RQ2: What is the gain obtained in the implementation of sequence alignment algorithms in FPGAs and GPUs? We aim to understand the benefits of implementing the sequence alignment algorithms in those hardware. It is expected to observe an improvement in the execution time, energy efficiency and memory consumption.

39.2.2 Research Execution

Three databases, IEEE, Scopus and ACM, were used as sources of publications for the later systematic mapping. Only

Table 39.1 Inclusion (IC) and Exclusion (EC) Criterias

ID	Description
IC-01	Publications with a algorithm implementation in FPGAs and GPUs can be selected.
IC-02	Publications with a mention of some improvement by using FPGAs and GPUs can be selected.
IC-03	Publications that propose different approaches for sequence alignment algorithm can be selected.
EC-01	Publications without the keywords in the title and/or abstract will not be selected.
EC-02	Publications with comparisons between algorithms will not be selected.
EC-03	Publications about non-accelerated hardware algorithms implementation will not be selected.
EC-04	Publications prior to 2015 will not be selected.

publications dating from 2015 were allowed in the searches. Then, the following search expression were elaborated and submitted to the three databases:

((*“hardware implementation” OR “hardware”*) AND (*“sequence alignment algorithm” OR “sequence alignment”*)) AND (*“FPGA” OR “GPU” OR “hardware description” OR “verilog” OR “VHDL”*)

The number of articles obtained by the search string in IEEE, Scopus and ACM were, respectively, 26, 54 and 97. After removing the duplicate occurrences, the total of manuscripts decreased from 177 to 137. Then, we elaborated a set of inclusion and exclusion criteria and we applied them to the abstracts of the articles. After this step, the number of papers decreased again to 29. Table 39.1 describes the inclusion and exclusion criteria used to refine the remaining manuscripts.

Finally, we applied one more filter to the remaining articles. We read them to verify if they belonged to the full context of our systematic mapping. Table 39.2 shows the 23 selected articles.

39.2.3 Data Extraction

After the articles selection for mapping, it was proceeded to the data extraction process. In this step, we sought to get the information to answer the proposed questions. The data extracted from each article were used to fill the following fields:

- title;
- publication year;
- authors;
- implemented algorithm or used as a basis;
- hardware used (FPGA or GPU).

Table 39.2 Selected research articles

Authors	Title
Gálvez, Sergio et al. (2016) [6]	Speeding-up bioinformatics algorithms with heterogeneous architectures: Highly heterogeneous Smith-Waterman (HHeterSW)
Ahmed, Nauman et al. (2015) [7]	Heterogeneous hardware/software acceleration of the BWA-MEM DNA alignment algorithm
Arram, James et al. (2016) [8]	Leveraging FPGAs for accelerating short read alignment
Bekbolat, Marzhan et al. (2019) [9]	HBLast: An open-source FPGA library for DNA sequencing acceleration
Chacón, Alejandro et al. (2014) [10]	Boosting the FM-index on the GPU: Effective techniques to mitigate random memory access
Chen, Nae-Chyun et al. (2018) [11]	A memory-efficient FM-index constructor for next-generation sequencing applications on FPGAs
Di Tucci, Lorenzo et al. (2018) [12]	A parallel, energy efficient hardware architecture for the merAligner on FPGA using chisel HCL
Fakirah, Maged et al. (2015) [13]	Accelerating Needleman-Wunsch global alignment algorithm with GPUs
Fei, Xia et al. (2017) [14]	FPGASW: Accelerating large-scale Smith-Waterman sequence alignment application with backtracking on FPGA linear systolic array
Gajda, Dominik et al. (2018) [15]	BioCircuit – A hardware based methodology for protein recognition
El-Wafa et al. (2016) [16]	Hardware acceleration of Smith-Waterman algorithm for short read DNA alignment using FPGA
Sitao Huang et al.(2017) [17]	Hardware acceleration of the pair-HMM algorithm for DNA variant calling
Ibrahim et al. (2016) [18]	Novel reconfigurable hardware accelerator for protein sequence alignment using Smith-Waterman algorithm
Jiang et al. (2018) [19]	CUDAMPF++: A proactive resource exhaustion scheme for accelerating homologous sequence search on CUDA-enabled GPU
Li et al. (2019) [20]	BLASTP-ACC: Parallel architecture and hardware accelerator design for BLASTbased protein sequence alignment
Liu et al. (2016) [21]	A customized many-core hardware acceleration platform for short read mapping problems using distributed memory interface with 3D-stacked architecture
Patrick et al. (2018) [22]	RNS Smith-Waterman accelerator based on the moduli set $2n, 2n-1, 2n-1-1$
Nurdin, D.et al. (2016) [23]	DNA sequence alignment: A review of hardware accelerators and A new core architecture
Seliem, Asmaa et al. (2016) [24]	Parallel Smith-Waterman algorithm hardware implementation for ancestors and offspring gene tracer
Waidyasooriya, Hasitha et al. (2016) [25]	Hardware-acceleration of short-read alignment based on the Burrows-Wheeler transform
Warris, Sven et al. (2018) [26]	pyPaSWAS: Python-based multi-core CPU and GPU sequence alignment
Yoshimi, Masato et al. (2016) [27]	Accelerating BLAST computation on an FPGA-enhanced PC cluster
Zeni, Alberto et al. (2019) [28]	circFA: a FPGA-based circular RNA aligner

39.3 Results

In this section, the results obtained after extracting the data of interest are discussed. It was identified different sequence alignment algorithms implemented in the selected articles. The Smith-Waterman algorithm is the most used one, whilst FPGA was the most used hardware. The following subsections details the results. First, the FPGAs and GPUs usage is presented. Then, the different sequence alignment algorithms implemented in those hardware are presented. Finally, the gains of those algorithms are analyzed.

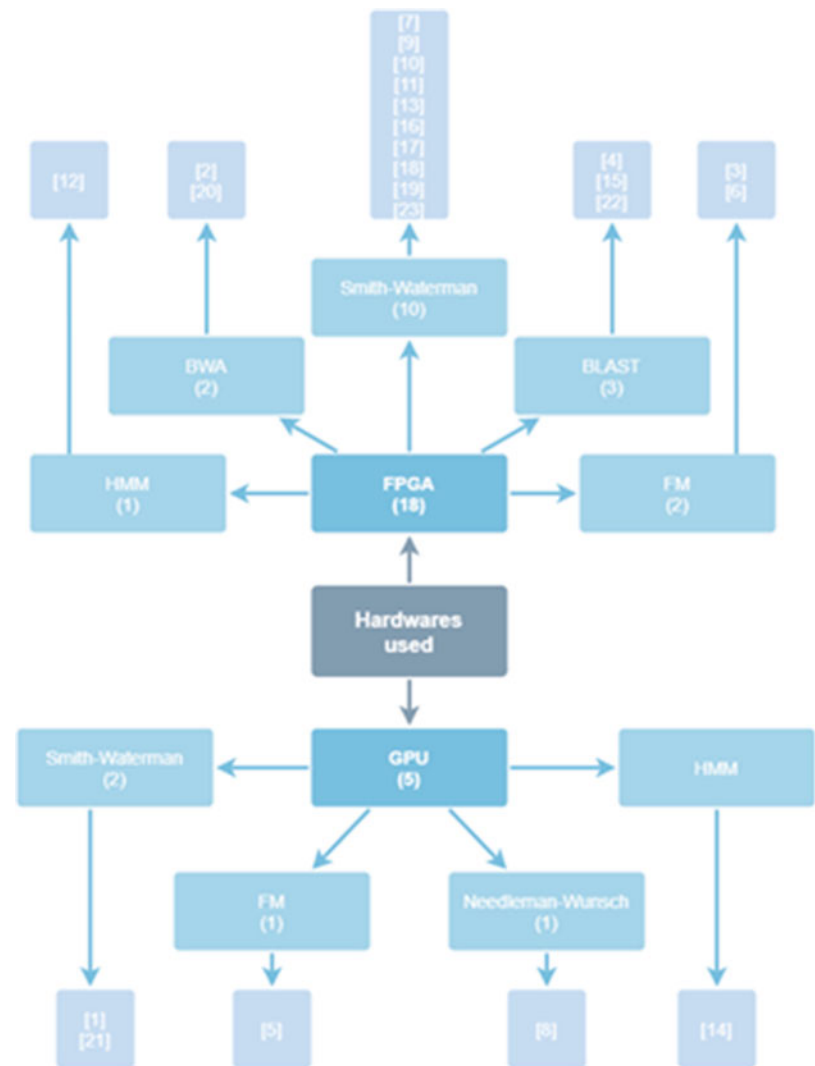
39.3.1 FPGAs and GPUs

As previously mentioned, the FPGA was the most used hardware, being used in 18 articles (approximately 78%). Thus, five articles used GPU as the chosen approach.

39.3.2 Implemented Algorithms

Figure 39.1 shows a conceptual map with the algorithms used for each platform and article. The data from the systematic mapping concerning the algorithm employed shows the number of articles that use each algorithm. Amongst the mapped

Fig. 39.1 Conceptual map of the algorithms used by GPU and FPGA based on the selected articles



works, we found 6 different sequence alignment algorithms. The Smith-Waterman algorithm was used in more than half of the selected works (12 times). Next, the BLAST algorithm [29] was used in 3 articles, followed by the BWA algorithm [30] used twice. The Needleman-Wunsch algorithm was used only once, and 3 works used FM-Index [31]. Additionally, [10] presents a special data structure to hasten searches on large data sets. Finally, two papers presented approaches based on HMM [32]. In [17], the Pair-HMM [33] was used, which has been derived from HMM. Although some of these sequence alignment algorithms have different functioning mechanisms and purposes, they can be implemented in FPGAs and GPUs.

39.3.3 Improved Performance and Energy Efficiency

We observed on the selected papers that the algorithm implementation choice in hardware occurs by using parallelism

efficiently. Since we have to process extensive data sets, the use of parallelism makes the execution time considerably shorter, which is a significant gain for the problems in this scope. We can observe an important aspect from the papers used in the mapping, the algorithm implementation choice in hardware occurs by using parallelism. Since we have to process extensive data sets, parallelism makes the execution time shorter, which is a significant gain for the problems in this subject.

This improvement in the execution time is a common result in the analyzed works. When compared to the version of the algorithm implemented in CPU, the article [8] obtained an execution time about 28 times faster in FPGA, while the results from [10] were 8 times faster in GPU. Besides that [14], implemented the Smith-Waterman between 3.6 and 25.2 times faster in FPGA when compared to the implementation in CPU. Consequently, we can infer that FPGA implementations are even faster than GPU implementations, and this may explain why the former were more common on the selected articles.

The implementation of the Smith-Waterman algorithm presented in [16] shows an execution time improvement in FPGA of 28.4 times when compared to the respective CPU algorithm. This work also concerns about memory usage during the algorithm execution in a parallel hardware, stating that FPGA implementation uses less memory than the sequential one.

The article [12] also observed gains in energy efficiency. By implementing the Smith-Waterman algorithm using FPGA, they obtained not only an execution time 7 times faster, but also an energy efficiency from 8 to 66 times higher.

We conclude that the mapped works had a significant execution time improvement using FPGAs and GPUs. This was an expected result, since those hardware provide parallelism for the sequence alignment algorithms, which considerably reduces their execution time. The energy gain obtained as a result in the article [12] is important to consider, since it shows how this approach of systematic mapping may lead to surprising and unexpected results.

39.4 Discussion

In this section, we discuss the two research questions used to guide the systematic review.

RQ1: Which sequence alignment algorithms have been implemented in FPGAs and GPUs in recent years? As shown in Sect. 39.2.3, we found several sequence alignment algorithms, but the Smith-Waterman is the most widely used in the analyzed works.

The Smith-Waterman algorithm derives from the Needleman-Wunsch algorithm, and both can be accelerated by hardware parallelism. Although the first aims the global alignment while the second aims the local alignment, both find the best answer for the sequence alignment. As argued in [34], the Smith-Waterman algorithm has a quadratic complexity, or $O(xy)$ when considering two sequences with sizes x and y . This implies a longer execution time for analyzing a large data set. In this way, the implementation of these algorithms in hardware becomes advantageous, since it takes use of the parallelism.

The BLAST algorithm has a low complexity but, unlike Smith-Waterman and Needleman-Wunsch algorithms, it does not guarantee the best solution to the sequence alignment problem. In fact, as said in [29], this algorithm is a variation of the Smith-Waterman algorithm itself that tries to obtain less time consumption by using an optimized model. As a consequence, it provides a less accurate result.

Thus, the choice of using the Smith-Waterman algorithm is a plausible choice because, despite it has a quadratic complexity when sped up using hardware, it guarantees to find

the best solution while still showing a decent improvement in execution time.

RQ2: What is the gain obtained in the implementation of sequence alignment algorithms in FPGAs and GPUs?

As specified in [7, 10, 12], the main benefit of this approach is the execution time.

In this way, large pairs of data strings executes in considerably faster, resulting in a sensible optimization of the information processing. As stated in Sect. 39.2.3, the parallelization and the processing power of the selected hardware were the most responsible for the observed performance increase.

Another gain mentioned is energy efficiency, as presented in [12]. This important feature was neglected in the others studies, and thus it arises as a research field for future works. It is possible to verify the advantages of using a given hardware regarding energy efficiency, or even the memory use, a result presented in [16].

39.5 Conclusion

In this work, the main sequence alignment algorithms implemented in hardware were identified, as well as the hardware used in the implementation. A systematic mapping was carried out to find the algorithms and hardwares used, in which 177 works were obtained. After filtering and analyzing the articles, a final number of 23 papers was analyzed. In general, the work contributed to a better understanding of the sequence alignment algorithms area, making it possible to have an overview of all the algorithms implemented in hardware in recent years and their respective gains when using this approach. Finally, it was possible to identify areas of study and research related to this environment.

References

1. B.S.C. Varma, K. Paul, M. Balakrishnan, *Architecture Exploration of FPGA Based Accelerators for Bioinformatics Applications* (Springer, Singapore, 2016)
2. A.Y. Zomaya et al., *Parallel Computing for Bioinformatics and Computational Biology* (Wiley Online Library, Hoboken, 2005)
3. S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**(3), 443–453 (1970)
4. T.F. Smith, M.S. Waterman, et al., Identification of common molecular subsequences. *J. Mol. Biol.* **147**(1), 195–197 (1981)
5. D. Tranfield, D. Denyer, P. Smart, Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *Br. J. Manag.* **14**(3), 207–222 (2003)
6. S. Galvez, A. Ferusic, F.J. Esteban, P. Hernandez, J.A. Caballero, G. Dorado, Speeding-up bioinformatics algorithms with hetero-

- geneous architectures: Highly heterogeneous Smith-Waterman (HHeterSW). *J. Comput. Biol.* **23**(10), 801–809 (2016)
7. N. Ahmed, V.-M. Sima, E. Houtgast, K. Bertels, Z. Al-Ars, Heterogeneous hardware/software acceleration of the BWA-MEM DNA alignment algorithm, in *2015 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, (IEEE, 2015), pp. 240–246
 8. J. Arram, T. Kaplan, W. Luk, P. Jiang, Leveraging FPGAs for accelerating short read alignment. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **14**(3), 668–677 (2016)
 9. M. Bekbolat, S. Kairatova, A. Shymyrbay, K. Vipin, HBLast: An open-source FPGA Library for DNA sequencing acceleration, in *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, (IEEE, 2019), pp. 79–82
 10. A. Chacon, S. Marco-Sola, A. Espinosa, P. Ribeca, J.C. Moure, Boosting the FM-index on the GPU: Effective techniques to mitigate random memory access. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **12**(5), 1048–1059 (2014)
 11. N.-C. Chen, Y.-C. Li, Y.-C. Lu, A memory-efficient FM-index constructor for next-generation sequencing applications on FPGAs, in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, (IEEE, 2018), pp. 1–4
 12. L. Di Tucci, D. Conficconi, A. Comodi, S. Hofmeyr, D. Donofrio, M.D. Santambrogio, A parallel, energy efficient hardware architecture for the merAligner on FPGA using Chisel HCL, in *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, (IEEE, 2018), pp. 214–217
 13. M. Fakirah, M.A. Shehab, Y. Jararweh, M. AlAyyoub, Accelerating Needleman-Wunsch global alignment algorithm with GPUs, in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, (IEEE, 2015), pp. 1–5
 14. X. Fei, Z. Dan, L. Lina, M. Xin, Z. Chunlei, FPGASW: Accelerating large-scale Smith–Waterman sequence alignment application with backtracking on FPGA linear systolic array. *Interdiscip. Sci. Comput. Life Sci.* **10**(1), 176–188 (2018)
 15. D. Gajda, A. Pulka, BioCircuit-a hardware based methodology for protein recognition, in *2018 International Conference on Signals and Electronic Systems (ICSES)*, (IEEE, 2018), pp. 289–294
 16. W. Abou El-Wafa, A.G. Seliem, H.F. Hamed, Hardware acceleration of Smith-Waterman algorithm for short read DNA alignment using FPGA, in *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, (IEEE, 2016), pp. 604–605
 17. S. Huang, G.J. Manikandan, A. Ramachandran, K. Rupnow, W.-M.W. Hwu, D. Chen, Hardware acceleration of the pair-HMM algorithm for DNA variant calling, in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, (2017), pp. 275–284
 18. A. Ibrahim, H. Elsimary, A. Aljumah, Novel reconfigurable hardware accelerator for protein sequence alignment using Smith-Waterman algorithm. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **99**(3), 683–690 (2016)
 19. H. Jiang, N. Ganesan, Y.-D. Yao, CUDAMPF++: A proactive resource exhaustion scheme for accelerating homologous sequence search on CUDA-enabled GPU. *IEEE Trans. Parallel Distrib. Syst.* **29**(10), 2206–2222 (2018)
 20. Y.-C. Li, Y.-C. Lu, BLASTP-ACC: Parallel architecture and hardware accelerator design for BLASTbased protein sequence alignment. *IEEE Trans. Biomed. Circuits Syst.* **13**(6), 1771–1782 (2019)
 21. P. Liu, A. Hemani, K. Paul, C. Weis, M. Jung, N. Wehn, A customized many-core hardware acceleration platform for short read mapping problems using distributed memory interface with 3D-stacked architecture. *J. Signal Process. Syst.* **87**(3), 327–341 (2017)
 22. P.K. Mensah, E.K. Bankas, M.M. Iddrisu, RNS Smith-Waterman Accelerator based on the moduli set $2^n, 2^{n-1}, 2^{n-1}-1$, in *2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST)*, (IEEE, 2018), pp. 1–8
 23. D. Nurdin, M. Isa, S. Goh, DNA sequence alignment: A review of hardware accelerators and a new core architecture, in *2016 3rd International Conference on Electronic Design (ICED)*, (IEEE, 2016), pp. 264–268
 24. A.G. Seliem, W. Abou El-Wafa, A. Galal, H.F. Hamed, Parallel Smith-Waterman algorithm hardware implementation for ancestors and offspring gene tracer, in *2016 World Symposium on Computer Applications & Research (WSCAR)*, (IEEE, 2016), pp. 116–121
 25. H.M. Waidyasooriya, M. Hariyama, Hardware acceleration of short-read alignment based on the Burrows-Wheeler transform. *IEEE Trans. Parallel Distrib. Syst.* **27**(5), 1358–1372 (2015)
 26. S. Warris, N.R.N. Timal, M. Kempenaar, A.M. Poortinga, H. van de Geest, A.L. Varbanescu, J.-P. Nap, pyPaSWAS: Python-based multi-core CPU and GPU sequence alignment. *PLoS One* **13**(1), e0190279 (2018)
 27. M. Yoshimi, C. Wu, T. Yoshinaga, Accelerating BLAST computation on an FPGA-enhanced PC cluster, in *2016 Fourth International Symposium on Computing and Networking (CANDAR)*, (IEEE, 2016), pp. 67–76
 28. A. Zeni, F. Peverelli, E. Cabri, L. Di Tucci, L. Cerina, M.D. Santambrogio, circFA: a FPGA-based circular RNA aligner, in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, (IEEE, 2019), pp. 1–4
 29. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**(3), 403–410 (1990)
 30. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009)
 31. P. Ferragina, G. Manzini, Opportunistic data structures with applications, in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, (IEEE, 2000), pp. 390–398
 32. S.R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**(10), e1002195 (2011)
 33. G. Stormo, Book review: Biological sequence analysis: Probabilistic models of proteins and nucleic acids Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison. *Q. Rev. Biol.* **75**, 09 (2000)
 34. S.K. Zahid, L. Hasan, A.A. Khan, S. Ullah, A novel structure of the Smith-Waterman algorithm for efficient sequence alignment, in *2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC)*, (IEEE, 2015), pp. 6–9

Hardware Logic Library and High-Level Logic Synthesizer Combining LOTOS and a Functional Programming Language

40

Katsumi Wasaki

Abstract

LOTOS is a formal description language for parallel computing models based on process algebra. There is an arithmetic element library, DILL, for describing logic circuits on LOTOS. However, due to the LOTOS description's rigor and the classical m4 macro language expansion method, full-scale circuit design has been complicated. This study aims to improve the hardware upper-level design's descriptiveness by constructing a new arithmetic element library using syntax sugar written in OCaml, a functional language called HDCaml, and converting it LOTOS code while using the features of a high-level description language. We have developed a LOTOS code generator and the comprehensive library, as OpenDill, which consists of primary logic gate elements (AND, OR, and selector). OCaml has various library construction features, such as class management, reusability, and data-driven dynamic generation of instances. In the future, we plan to build a library of complex circuit elements (fast adder, ALU) in HDCaml description.

Keywords

Upstream hardware compiler · Code generation · Logic synthesizer · Formal specification language · LOTOS · OCaml · Formal method · m4 macro · Logic library · OpenDill

40.1 Introduction

Formal verification has been studied to check whether a system works properly according to its specifications. Formal verification is a method to logically and algebraically describe the system configuration and check the system's behavior described by a computer exhaustively. For instance, LOTOS, VDM-SL, and the Z language have been developed as a formal description language for system design [1–3]. A Petri net is one of several mathematical modeling languages for the description and verification of distributed systems [4]. On the other hand, there are many studies on mathematical models of logic circuits, and libraries for formal verification are being enhanced [5].

LOTOS is a formal description language with abstract data types and Communicating Sequential Processes [6] capabilities for describing specifications of asynchronous systems, protocols, etc. ISO has standardized the LOTOS syntax. An m4 macro library called DILL [7] has been developed to perform dynamic verification of logic gate circuits on LOTOS. This logic library describes the logic gates' behavior as a white-box (describing all the changes in circuit connections and internal states) and a black box (describing the device's behavior). Each specification starts from a primary logic gate such as AND, OR, and XOR with 2 to 8 inputs. The behavioral specifications of complex circuit elements, such as adders, latches, data selectors, and memory elements, are organized sequentially as a hierarchical process. However, in the case of DILL circuit design, the synchronization of each circuit element's input and output lines had to be strictly described in the LOTOS language. Using classical m4 macro language preprocessor and in-lay expansion method makes it difficult to describe the full-scale circuits to worsen.

K. Wasaki (✉)
Faculty of Engineering, Shinshu University, Nagano, Japan
e-mail: wasaki@cs.shinshu-u.ac.jp

In this study, we construct a new arithmetic element library, OpenDill [8], similar to DILL, and convert HDCaml [9] code to LOTOS code using a syntax sugar for hardware-level design written in Ocaml [10] which is migrating to a meta hardware description language [11]. With this ingenuity, we have succeeded in improving the language's descriptiveness while utilizing the features of high-level description languages. We have developed a LOTOS code generator for a library consisting of essential logic elements (basic gates, MUX, and selectors). Functions such as numeric operations for multi-bit input and storing the results in synchronous registers can also be described using the definitions of various higher-level functional elements in the DILL library. This study's advantage is that the synthesis and code generation of logic circuits and code generation can be done directly from the high level of abstraction using higher-order functions of OCaml.

40.2 Hardware Logic Library: OpenDill

There is an existing logic circuit library DILL written in LOTOS. This library is a set of definitions of abstract data types and processes for describing logic circuits' behavior and abstracting the input/output relationships of essential gate elements. Existing research has been building libraries that describe the operating processes of arithmetic and functional devices such as adders and flip-flops.

In this study, a method for generating LOTOS code from circuit descriptions written in HDCaml is investigated, and the circuit library DILL is extended to apply to the high-level design and maintained as OpenDill. Specifically, we aimed to generate the LOTOS code corresponding to the RTL description of HDCaml using the complex circuit definitions of the DILL logic circuit library.

40.2.1 Defining Abstract Data Types

In DILL library, data types and various logic operations functions are defined as abstract data types representing logic

gate circuits on LOTOS. The data types are defined as three values (low, high, and indefinite) that can be taken by a 1-bit circuit line and a type with a binary operator name as a value. The definition of the abstract data type is described in the form ACT ONE [12].

Besides, a function called 'Apply' is defined as a method of abstraction for the internal arithmetic process described below. The function is constructed so that by giving the target operator and the operator the name to be operated on as arguments, it calls the target operation internally and returns the output value. Figure 40.1 shows the definition of the Apply function for the three inputs. When the Apply function is called together with the three inputs, it is defined to return the various binary operations of type Bit defined in advance. Also, if a unary operation (e.g., NOT) is specified and called despite the input, the output value is defined to be treated as undefined.

40.2.2 Abstract Model of Logic Gates

In DILL library, gate devices are described by giving functions to process descriptions abstracted from input-output relations only, without describing the behavior of devices such as AND, OR, etc. Figure 40.2 shows a description of the abstraction process (Logic3) of a three-input gate device. The internal logic is nondeterministically selective and replaces the internal values at the input values' change points. For the input, the Apply function determines the output value and passes the value to the output line only when the output value changes.

DILL includes essential gate elements and behavioral descriptions of functional elements such as adders and flip-flops. Each device is described as either a white-box type (configured as an asynchronous circuit connection) or a black-box type (configured as an external input/output relationship). For example, the white-box definitions of the 2-2 half-adders and 3-2 full-adders in DILL are written, as shown in Fig. 40.3. The hierarchical synthesis of essential circuit

Fig. 40.1 Part of the Apply function definition (LOTOS, OpenDill.lib in ACT ONE)

```

type BitOp3 is BitOp2
  opns
    Apply : BitOp, Bit, Bit, Bit -> Bit
  eqns
    forall b1, b2, b3 : Bit, bop : BitOp
  ofsort Bit
    IsUnary(bop) =>
      Apply(bop, b1, b2, b3) = X; (* type mismatch *)
  ofsort Bit
    Apply(and, b1, b2, b3) = (b1 gand b2) gand b3;
    Apply(nand, b1, b2, b3) = gnot ((b1 gand b2) gand b3);
    Apply(or, b1, b2, b3) = (b1 gor b2) gor b3;
    Apply(nor, b1, b2, b3) = gnot ((b1 gor b2) gor b3);
    Apply(xor, b1, b2, b3) = (b1 gxor b2) gxor b3;
    Apply(xnor, b1, b2, b3) = gnot((b1 gxor b2) gxor b3);
  endtype (* BitOp3 *)

```


Fig. 40.2 Abstraction description of the 3-input generalized logic gate in DILL (LOTOS and m4, dill_gate.m4)

```

define(Logic3_Decl,`declare(`$0',`
process Logic3 [Ip1, Ip2, Ip3, Op] (BOp : BitOp) : noexit
:= Logic3Aux [Ip1, Ip2, Ip3, Op] (BOp, X of Bit, X of
Bit, X of Bit, X of Bit)
where
process Logic3Aux [Ip1, Ip2, Ip3, Op]
(BOp : BitOp, BIn1, BIn2, BIn3, BOut : Bit) : noexit
:=
Ip1 ? BIn1New : Bit;
Logic3Aux [Ip1, Ip2, Ip3, Op]
(BOp, BIn1New, BIn2, BIn3, BOut)
[]
Ip2 ? BIn2New : Bit;
Logic3Aux [Ip1, Ip2, Ip3, Op]
(BOp, BIn1, BIn2New, BIn3, BOut)
[]
Ip3 ? BIn3New : Bit;
Logic3Aux [Ip1, Ip2, Ip3, Op]
(BOp, BIn1, BIn2, BIn3New, BOut)
[]
(
let BOutNew:Bit = Apply(BOp, BIn1, BIn2, BIn3) in
[(BOutNew eq X of Bit)and(BOut eq X of Bit)] ->
Op ? BOutNew : Bit [BOutNew ne X of Bit];
Logic3Aux [Ip1, Ip2, Ip3, Op]
(BOp, BIn1, BIn2, BIn3, BOutNew)
[]
[(BOutNew ne X of Bit)and(BOutNew ne BOut)] ->
Op ! BOutNew;
Logic3Aux [Ip1, Ip2, Ip3, Op]
(BOp, BIn1, BIn2, BIn3, BOutNew)
)
endproc (* Logic3Aux *)
endproc (* Logic3 *)
')')

```

Fig. 40.3 Definitions of the 2-2 half-adder and 3-2 full-adder (white-box, dill_adder.m4)

```

define(HalfAdder_Decl,`declare(`$0',`Xor2_Decl`'And2_Decl
process HalfAdder [A, B, S, C] : noexit :=
Xor2 [A, B, S]
|[A, B]|
And2 [A, B, C]
endproc (* HalfAdder *)
')')

define(FullAdder_Decl,`declare(`$0',`Or2_Decl`'HalfAdder_Decl
process FullAdder [A, B, Cin, S, Cout] : noexit :=
hide Sint, Cint0, Cint1 in
(
HalfAdder [A, Sint, S, Cint0]
|[Sint]|
HalfAdder [B, Cin, Sint, Cint1]
)
|[Cint0, Cint1]|
Or2 [Cint0, Cint1, Cout]
endproc (* FullAdder *)
')')

```

Fig. 40.4 Definition of a binary counter with 4-bit presets (white-box, `dill_counter.m4`)

```
define(Bi_Counter4_Reset_Decl,
`declare(`$0',`JKFlipFlop_PreClr_Decl`'Nand2_Decl`
'One_Decl
process Bi_Counter4_Reset [Q4, R1, R2, MWire(4, Q)]
:noexit :=
hide r1r2, MWire(4, `J, K, Pre, Qbar') in
Nand2 [R1, R2, r1r2]
|[r1r2]|
MComp(4, `r1r2=, Q',
`JKFlipFlop_PreClr [J, K, Pre, r1r2=, Q+, Q, Qbar]')
|[MWire(4, `J, K, Pre')]|
(MComp(4, One [J]) ||| MComp(4, One [K])
||| MComp(4, One [Pre]))
endproc (* Bi_Counter4_Reset *)
')
```

Table 40.1 Existence of various component libraries in DILL

Element type	White-box	Black-box
Adder	o	o
Encoder/decoder	o	o
Comparator	o	o
Parity checker/generator	o	o
Multiplexer/demultiplexer	o	o
Latch	o	o
Flip-flop	o	o
Counter	o	x
Register	o	x
Memory	o	x
Filter	x	x
Parallel multiplexer	x	x

elements describes the internal structure, and the signal lines shared by each element are explicitly described.

On the other hand, the definitions of various circuit elements in the DILL library are shown in Table 40.1. The elements up to the memory are defined in the white-box type, but no higher-level circuit design is defined. For example, since there is no library of filters and parallel multipliers, we have extended the definition of these high-level devices in this study.

40.2.3 Example Description of a Repetitive Circuit Configuration

DILL has `m4` macros for describing repetitive structures called `Mcomp` and `MWire`, which define multiple devices and connection lines in succession. We thus can define circuits with recursive structures by improving the description.

For example, the structure of a 4-bit wide binary counter with presets is shown in Fig. 40.4. The element's internal structure has a recursive structure in which an iterative macro `Mcomp` generates four stages of JK flip-flops and connects the digit-up output of the previous stage to the upper module's input, respectively. Also, the definition of the

```
open Hdcaml;;
open Design;;
open Lotos;;
```

Fig. 40.5 Declarations of library packages using HDCaml and OpenDill

connection lines to the outside of the process and the hidden connection lines are repeatedly duplicated and described by the `MWire` macro.

40.3 High-Level Design Description in HDCaml

40.3.1 Overview

HDCaml is a syntax sugar for describing hardware circuit design in OCaml, a functional language, and enables circuit design in RTL (Register Transfer Level). Besides, it is written in OCaml's grammar, so it is possible to construct a circuit using the functions in OCaml. For the described circuits, the circuit models at the data flow level of Verilog [13] and VHDL [14], and the SystemC [15] circuit model can be output. In this study, a LOTOS code generator is implemented on OCaml.

Various functions of HDCaml can be used by calling the following modules at the beginning of the program that describes the circuit (the line of `open Lotos` is the `m4` program file of the code generator we created) in Fig. 40.5.

40.3.2 Example of a High-Level Design Description

Boolean expressions are described by a series of binary operators that have been extended by HDCaml to describe logical operations on OCaml. The operators defined include bitwise operations, addition and subtraction, multiplication, and shift operations.

Fig. 40.6 Example of a carry lookahead adder description in HDCaml (m4 program, excerpt, 4-bit width)

```

...
let rec carry_lookahead_generator g_l p_l terms =

  let term g_i p_i term_list =
    g_i :: List.map (fun t -> t &: p_i) term_list
  in
  let rec expression terms =
    match terms with
    | [] -> raise Empty_List
    | [t] -> t
    | t::rest -> t |>: expression rest
  in
  match g_l,p_l with
  | [],[] -> []
  | g_i::g_r,p_i::p_r ->
    let t_l = term g_i p_i terms in
    (expression terms)
    :: carry_lookahead_generator g_r p_r t_l
  | _ -> raise Input_Fail
;;

let carry_lookahead_sum_part a_l b_l c_i =

  let s_i a_i b_i c_i = a_i ^>: b_i ^>: c_i in
  let rec g_p_generate a_l b_l =
    let c_l_sum a_i b_i =
      let g_i = a_i &: b_i in
      let p_i = a_i |>: b_i in
      g_i,p_i
    in
    match a_l,b_l with
    | [],[] -> [],[]
    | a_i::a_r,b_i::b_r ->
      let g_i,p_i = c_l_sum a_i b_i in
      let g_l,p_l = g_p_generate a_r b_r in
      g_i::g_l,p_i::p_l
    | _ -> raise Input_Fail
  in
  let g_l,p_l = g_p_generate a_l b_l in
  let c_l = carry_lookahead_generator g_l p_l [c_i] in

  List.map2 (fun x y -> x y)
  (List.map2 (fun x y -> x y)
   (List.map (fun x -> s_i x) a_l) b_l) c_l
;;
...
list_output (carry_lookahead_sum_part a_l b_l c0);
...

```

HDCaml defines a data type called circuit type to interpret the described logic circuit structure and use it in the object code output. When compiling, the circuit configuration is stored in a circuit variable, and when generating the code, it searches inside the variable and generates a string of the corresponding object code.

In the following, we describe an example of a high-level design description by HDCaml. Figure 40.6 shows an example of carry-lookahead adder written by HDCaml. Figure 40.7 shows the connection diagram after circuit synthesis.

The i -th ($i > 0$) carry value of the carry-lookahead adder is defined as follows.

$$\begin{aligned}
 c_i &= g_{i-1} + c_{i-1}p_{i-1} \\
 &= g_{i-1} + (g_{i-2} + c_{i-2}p_{i-2})p_{i-1} \\
 &= g_{i-1} + (g_{i-2} + (g_{i-3} + c_{i-3}p_{i-3})p_{i-2})p_{i-1}
 \end{aligned}$$

This expression is recursive for c_i and can be easily implemented as an OCaml function. The HDCaml code that corresponds to the above formula is part of the carry_lookahead_generator function. This function takes a pre-generated list of g_i and p_i and a list (terms) that holds each term on the above formula's right-hand side as arguments. Then, it constructs c_i ($i = 1, 2, 3, \dots$) recursively and passes the list of operations stored in each stage as the return value,

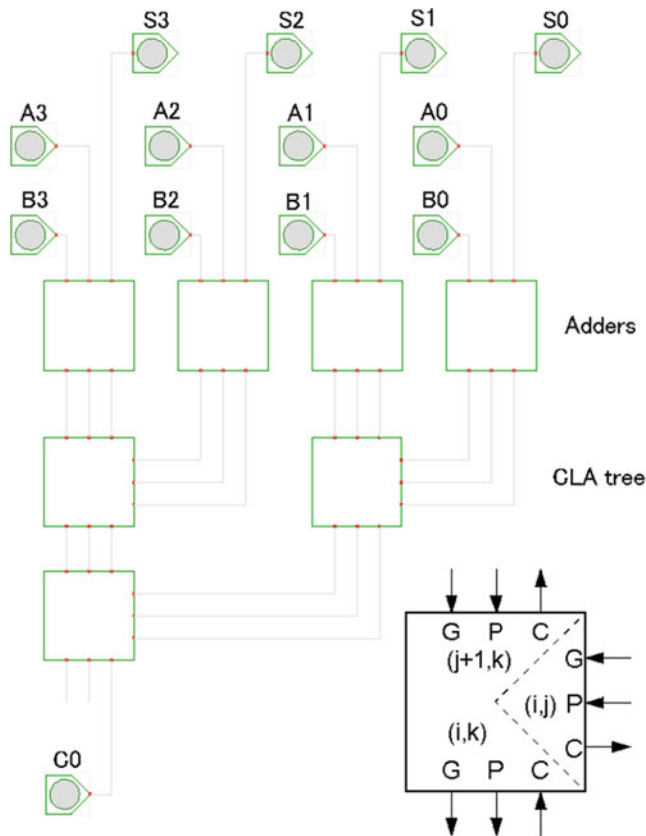


Fig. 40.7 Connection diagram of a carry lookahead adder after circuit synthesis (width = 4bit)

and passes a list containing each stage's operations as the return value.

The `carry_lookahead_sum_part` function generates a carry anticipation list internally with a list of input lines (`a_l`, `b_l`) and a carry input from the bottom row (`c_i`) as arguments. The final output line list is then derived by adding each element of the input line list and the carry operation list with the `s_i` function. We constructed a high-level design of the above n -bit arithmetic unit and migrated it as OpenDill.

40.4 LOTOS Code Generation for the OpenDill Library

We designed and tested a prototype LOTOS code generator based on the OpenDill library, extending the existing VHDL and Verilog code generators based on the DILL library.

40.4.1 LOTOS Code Generation Procedure

[Step-1] Defining Internal Connecting Lines and Type Resolution:

In LOTOS, when connecting multiple devices in a process,

it is necessary to connect and hide internal signal lines that do not appear in the operation's HDCaml description. For this reason, it is necessary to pre-describe the existence of the circuit lines to be used for the connections at the front of the process definition. In LOTOS's process generation, we search for the points connected by input/output between operators from the circuit types (structures) obtained by HDCaml parsing. It then generates internal synchronization event labels and lists the correspondence with the operators. Then, the internal signal lines are defined by calling the list at the time of process creation.

[Step-2] Adapting to RTL Circuit Description:

In HDCaml, addition and multiplication are defined as operators, and multi-bit operations are possible. However, in the circuit description by OpenDill, it is necessary to reduce the circuit representation to a bitwise operation by combining gate elements. Therefore, it is challenging to generate codes directly as n -bit operations such as n -bit addition and multiplication. Therefore, we use various logic operations modules with n -bit repetition structures defined in the OpenDill library described above section. By doing so, it is embedded in the LOTOS code generated from the circuit configuration and description of HDCaml, which corresponds to the description in RTL of HDCaml.

[Step-3] Consequences to Process Synchronization:

To describe a logic circuit with HDCaml, we only need to describe the definition of operations in RTL. However, LOTOS's definition requires each arithmetic process to explicitly describe synchronization between the connected mutual gates. Therefore, we use the correspondence list between the circuit structure, which stores the circuits' connection relations, and the arithmetic elements and internal synchronization lines. Specifically, when the LOTOS code is output for an operator process, the next operator process's connection is checked and output as a synchronization line. This strategy results in the synchronization of each arithmetic process.

40.4.2 High-Level Design, Synthesis, and LOTOS Code Generation

We developed a LOTOS code generator for HDCaml that has the above features. To evaluate the code generator, we performed a high-level trial using OpenDill for HDCaml high-level designs. Figure 40.8 shows an example of describing an n -bit parallel multiplier using addition and multiplication operators. Figure 40.9 shows the connection diagram after circuit synthesis.

The circuit configuration is as follows:

Fig. 40.8 Example description of an n-bit parallel multiplier using HDCaml (m4 program, excerpt, n-bit width)

```

...
let rec ripple_adder x_l y_l c_in =
  match x_l,y_l with
  | [],[] -> []
  | x::x_r,y::y_r ->
    let p_out,c_out = full_adder x y c_in in
    p_out :: ripple_adder x_r y_r c_out
  | _ -> raise Input_Fail
;;

let multiplier x_l y_l =
  let x_len = List.length x_l in
  let rec ini_list len =
    if len=0 then [] else (zero 1) :: ini_list (len-1)
  in
  let rec x_loop x_l y p_l c_l =
    match x_l with
    | [] -> [zero 1],[]
    | x::x_r ->
      let bit_mul = x & y in
      let p_out,c_out =
        full_adder bit_mul(List.hd p_l) (List.hd c_l)
      in
      let p_r,c_r =
        x_loop x_r y (List.tl p_l) (List.tl c_l) in
      p_out::p_r,c_out::c_r
  in
  let rec y_loop x_l y_l p_l c_l =
    match y_l with
    | [] -> ripple_adder p_l c_l (zero 1)
    | y::y_r ->
      let p_new,c_new = x_loop x_l y p_l c_l in
      (List.hd p_new)
      ::(y_loop x_l y_r (List.tl p_new) c_new)
  in
  y_loop x_l y_l (ini_list x_len) (ini_list x_len)
;;
...
list_output (multiplier [a1] [b1]);
...

```

Fig. 40.9 Connection diagram of a parallel multiplier after circuit synthesis (width = 2bit)

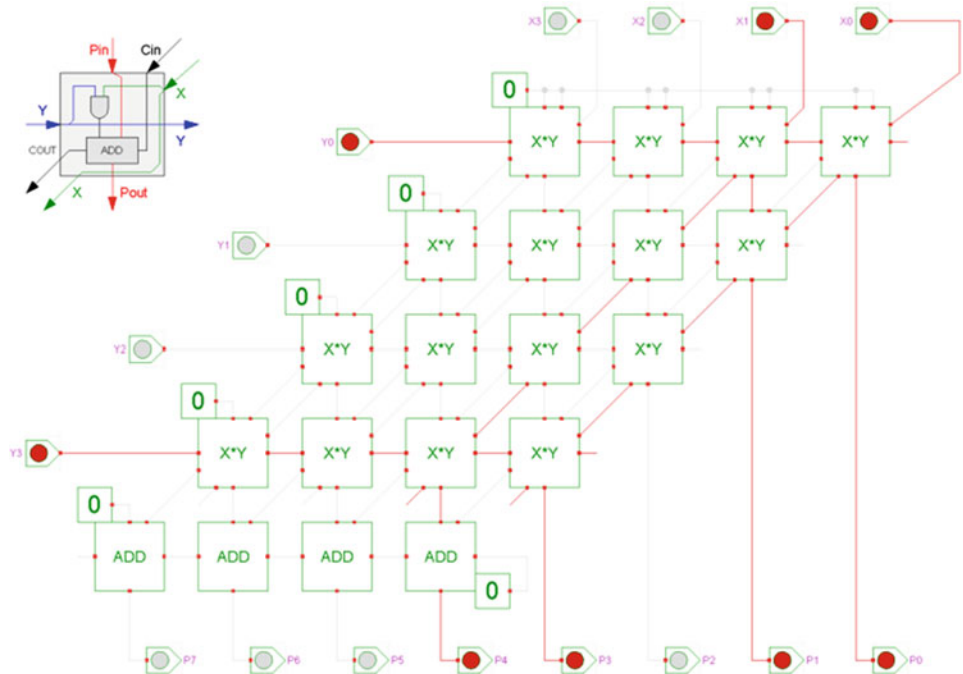


Fig. 40.10 Example of the output of the LOTOS code for the parallel multiplier (excerpt, width = 2bit)

```

...
process multiplier [ b1, b2, a1, a2,
                    output_3, output_2, output_1, output_0 ] : noexit :=
...
(Xor2 [n_55, n_69, output_3] |[n_55, n_69]|)
(Xor2 [n_52, n_61, n_69] |[n_52, n_61]|)
(Or2 [n_67, n_64, n_68] |[n_67, n_64]|)
(Or2 [n_66, n_65, n_67] |[n_66, n_65]|)
(And2 [n_55, n_52, n_66] |[n_55, n_52]|)
(And2 [n_52, n_61, n_65] |[n_52, n_61]|)
(And2 [n_61, n_55, n_64]
|[n_61]|)
( Xor2 [n_54, n_62, output_2] |[n_54, n_62]|)
(Xor2 [n_44, n_56, n_62] |[n_44, n_56]|)
(Or2 [n_60, n_57, n_61] |[n_60, n_57]|)
(Or2 [n_59, n_58, n_60] |[n_59, n_58]|)
(And2 [n_54, n_44, n_59] |[n_54, n_44]|)
(And2 [n_44, n_56, n_58] |[n_44, n_56]|)
(And2 [n_56, n_54, n_57] |[n_54]|)
(Xor2 [n_47, n_53, n_54] |[n_47, n_53]|)
(Xor2 [n_38, n_35, n_53] |[n_38, n_35]|)
(Or2 [n_51, n_48, n_52] |[n_51, n_48]|)
(Or2 [n_50, n_49, n_51] |[n_50, n_49]|)
(And2 [n_47, n_38, n_50] |[n_47, n_38]|)
(And2 [n_38, n_35, n_49] |[n_35]|)
(And2 [n_35, n_47, n_48] |[n_35, n_47]|)
(And2 [a2, b2, n_47]
|[a2, b2]|)
(Xor2 [n_39, n_45, output_1] |[n_39, n_45]|)
(Xor2 [n_37, n_27, n_45] |[n_37, n_27]|)
(Or2 [n_43, n_40, n_44] |[n_43, n_40]|)
(Or2 [n_42, n_41, n_43] |[n_42, n_41]|)
(And2 [n_39, n_37, n_42] |[n_39, n_37]|)
(And2 [n_37, n_27, n_41] |[n_37, n_27]|)
(And2 [n_27, n_39, n_40] |[n_27, n_39]|)
(And2 [a1, b2, n_39] |[a1]|)
(Xor2 [n_30, n_36, n_37] |[n_30, n_36]|)
(Xor2 [n_20, n_18, n_36] |[n_20, n_18]|)
(Or2 [n_34, n_31, n_35] |[n_34, n_31]|)
(Or2 [n_33, n_32, n_34] |[n_33, n_32]|)
(And2 [n_30, n_20, n_33] |[n_30, n_20]|)
(And2 [n_20, n_18, n_32] |[n_18]|)
(And2 [n_18, n_30, n_31] |[n_30]|)
(And2 [a2, b1, n_30]
|[b1]|)
(Xor2 [n_22, n_28, output_0] |[n_22, n_28]|)
(Xor2 [n_21, n_19, n_28] |[n_21, n_19]|)
(Or2 [n_26, n_23, n_27] |[n_26, n_23]|)
(Or2 [n_25, n_24, n_26] |[n_25, n_24]|)
(And2 [n_22, n_21, n_25] |[n_22, n_21]|)
(And2 [n_21, n_19, n_24] |[n_19]|)
(And2 [n_19, n_22, n_23] |[n_22]|)
And2 [a1, b1, n_22]
))))))))))))))))))))))))))))))))))))))))))))))
endproc
...

```

1. Prepare a recursive function (x_loop , y_loop) that performs the calculation configuration for each of the inputs x and y .
2. When a list of input y is read in one step, replicate the number of calculation modules for each input x (a double recursive structure).
3. Connect an adder from the replicated module's output line list (Fig. 40.8) that is finally obtained as the return value.

In the recursive module definition, the first element in the list of output p generated at each stage of y_loop is the output line that represents the result of the multiplication. For that reason, before calling the y_loop function for the next input

y, a list of entries is added to the list (`List.hd p_new:: (y_loop ...)`) as the return value of `y_loop`.

For the last adder in the bottom row of the circuit module, since the output sum of each adder represents the result of multiplication of the upper bits, every time all the adders are duplicated, they are successively added to the output list (`p_out:: ripple_adder ...`).

Finally, Fig. 40.10 shows the high-level synthesis results using OpenDill for the HDCaml high-level design shown in Fig. 40.8. This LOTOS code's output is obtained by compiling a 2-bit wide input list for a high-level description of an n-bit parallel multiplier.

In LOTOS, sub-processes are defined as nested (`process ... where ... endproc`) in the behavioral description. For this reason, when converting the code from HDCaml, the LOTOS code generation is delimited by the assignment unit of the called higher-order function, and the assignment destination is defined as a single process within the calling process, forming a nested structure. In the output example in Fig. 40.10, the circuit's internal configuration is generated in a scalable for each output line. Also, it was confirmed that the logic elements were connected regularly.

40.5 Conclusions and Future Work

The method of generating LOTOS code from circuit description written in HDCaml was studied, and the circuit library DILL was extended to apply to the high-level design and maintained as OpenDill. Specifically, the complex circuit definitions in the DILL logic circuit library are used to generate the LOTOS code corresponding to the RTL of HDCaml.

In the future, we will expand the functionality of the OpenDill library so that the current circuit description part can be hierarchically configured by using HDCaml only. OCaml has various library construction features, such as class management, reusability, and data-driven dynamic instance creation. Therefore, we would like to construct a

library of complex circuit elements (pipeline arithmetic [5], ALU) in HDCaml description using these features.

Acknowledgment This work was partially supported by JSPS KAKENHI Grant Number 19 K11821.

References

1. ISO/IEC 8807: Information Processing System, Open Systems Interconnection, LOTOS – A formal description technique based on the temporal ordering of observational behaviour (1989)
2. ISO/IEC 13817-1: Vienna Development Method – Specification Language – Part I: Base language (1996)
3. ISO/IEC 13568: Information technology, Z formal specification notation, Syntax, type system and semantics (2002)
4. Y. Harie, K. Wasaki, A Petri Net design and verification platform based on the scalable and parallel architecture: HiPS, in *Proceedings of the 14th International Conference on Information Technology – New Generations (ITNG2017), Advances in Intelligent Systems and Computing*, vol. 558, (Springer, 2017), pp. 265–273
5. K. Wasaki, Stability of the 7-3 compressor circuit for Wallace Tree. Part I. Formaliz. Math. **28**(1), 65–77 (2020)
6. C.A.R. Hoare, Communicating sequential processes. Commun. ACM **21**(8), 666–677 (1978)
7. J. He, K.J. Turner, Extended DILL: Digital Logic in LOTOS, Technical Report CSM-142, University of Stirling (1999)
8. Y. Kuwashima, K. Wasaki, Code generation and logic circuit library in formal specification language LOTOS using a high-level hardware design HDCaml, in *Proceedings of the 8th IPSJ Forum on Information Technology Conference (FIT)*, (2009), pp. 515–518
9. Karl Flicker: HDCaml improvements. Available at <http://karl-flicker.at/hdcaml/>
10. INRIA, France: OCaml. Available at <http://caml.inria.fr/ocaml/>
11. S. Nishida, K. Wasaki, Retargetable netlists generation and structural synthesis based on a meta hardware description language : Melasy+, in *Proceedings of the 9th International Conference on Information Technology: New Generations (ITNG2012)*, 827–830, *IEEE Computer Society Conference Proceedings*, (2012)
12. H. Ehrig, B. Mahr, *Fundamentals of Algebraic Specification, Part 1* (Springer Verlag, Berlin, 1985)
13. D.E. Thomas, P. Moorby, *The Verilog Hardware Description Language* (Kluwer Academic Publishers, 1991)
14. VHDL (VHSIC Hardware Description Language): Available at <http://vhdl.org/>
15. System C: Available at <http://www.systemc.org/>

Aavaas Gajurel, Sushil J. Louis, Rui Wu, Lee Barford,
and Frederick C. Harris Jr.

Abstract

In this paper, we use graphics processing units (GPU) to accelerate sparse and arbitrary structured neural networks. Sparse networks have nodes in the network that are not fully connected with nodes in preceding and following layers, and arbitrary structure neural networks have different number of nodes in each layers. Sparse Neural networks with arbitrary structures are generally created in the processes like neural network pruning and evolutionary machine learning strategies. We show that we can gain significant speedup for full activation of such neural networks using graphical processing units. We do a preprocessing step to determine dependency groups for all the nodes in a network, and use that information to guide the progression of activation in the neural network. Then we compute activation for each nodes in its own separate thread in the GPU, which allows for massive parallelization. We use CUDA framework to implement our approach and compare the results of sequential and GPU implementations. Our results show that the activation of sparse neural networks lends very well to GPU accel-

ation and can help speed up machine learning strategies which generate such networks or other processes that have similar structure.

Keywords

GPU · Neural networks · CUDA · Graph processing

41.1 Introduction

Artificial neural networks, first proposed by McCulloch and Pitts in 1943 [1], are universal function approximators loosely based on biological neural networks. Neural networks with back propagation [2] is a robust method in machine learning. Artificial neural networks (ANN) have interconnected nodes that are separated into three types—inputs, outputs and hidden nodes. Inputs nodes are sensor nodes that take in values from outside the system, output nodes are the nodes that produce the answers from the network and hidden nodes are the nodes which lie in the information propagation path of the neural network. Each node is activated based on the nodes from which it has incoming connections, and the activation is calculated by weighting all the incoming node values with corresponding connection weight and summing all the values. The sum is then thresholded for the final activation. Generally the sum is passed through sigmoid function to constrain it within -1 and $+1$ values.

$$\text{sigmoid}(x) = (1/(1 + e^{-4.97*x}))$$

Conventionally, neural networks have structure which consists of nodes cleanly segmented into layers as shown in Fig. 41.1 with incoming nodes shown in red and outgoing

A. Gajurel · S. J. Louis · F. C. Harris Jr. (✉)
Department of Computer Science and Engineering, University of
Nevada, Reno, Reno, NV, USA
e-mail: avs@nevada.unr.edu; sushil@cse.unr.edu;
fred.harris@cse.unr.edu

R. Wu
Department of Computer Science, East Carolina University,
Greenville, NC, USA
e-mail: wur18@ecu.edu

L. Barford
Department of Computer Science and Engineering, University of
Nevada, Reno, Reno, NV, USA

Keysight Laboratories, Keysight Technologies, Reno, NV, USA
e-mail: lee.barford@ieee.org

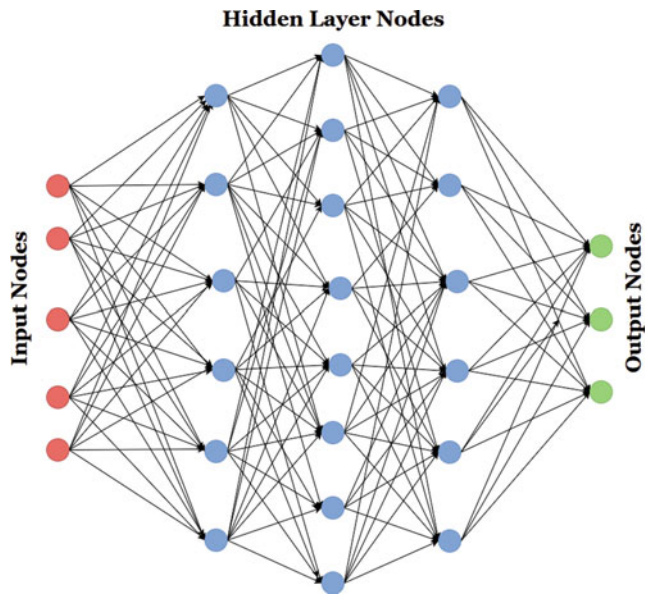


Fig. 41.1 Representation of conventional feed forward neural network with input, hidden and output layers

nodes shown in green. Nodes in each layer are not connected with each other and are fully connected with the nodes in preceding layer and the following layer. This structure lends very well to vectorization where each layer is represented by matrices of weights and the propagation and activation can be calculated with the multiplication of the layer matrices. This property is utilized for GPU acceleration of such neural network. As GPUs are well suited for large in matrix multiplications, such neural networks have seen large speedups, even for huge networks [3].

This paper is concerned with the activation of sparse and arbitrary structured neural networks. Neurons in sparse neural networks do not have full connection with nodes from preceding layer and following layer. Neural networks have arbitrary structure when nodes from sparse networks are also pruned. In such case, such networks cannot be cleanly separated into layers, i.e. they are not fully connected and can have incoming and outgoing links to any node in the graph as shown in Fig. 41.2. Sparse networks are a subset of arbitrary structured neural networks (ASNNs) and are generated by neural network pruning algorithms [4,5]. ASNNs can be generated by pruning both connections and nodes from fully trained dense networks. They are also created by rule based network structure generators, and some of which are applied in machine learning to generate networks best fit for a given problem. Neural Evolution of Augmenting Topologies (NEAT) [6] with direct encoding, HyperNEAT [7] which uses generative encoding, and grammar based substitution and bi-directional growth encoding [8] are some of the few processes which can generate ASNNs. With the start of application of neuro evolution to deep neural networks [9], full propagation of these network will take significant portion

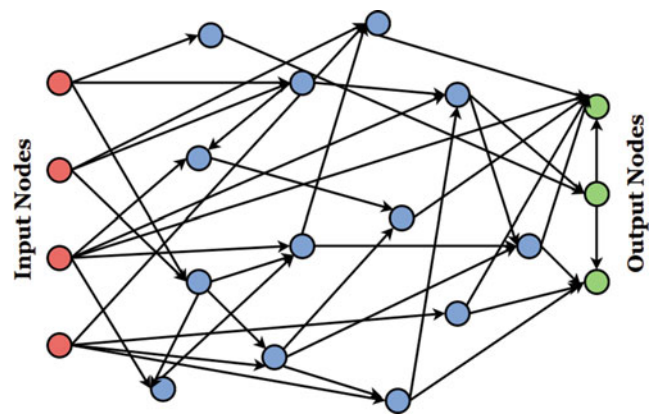


Fig. 41.2 Representation of neural network with arbitrary structure

of compute time of total program running time. GPU acceleration methodologies used for conventional NN will not work with ASNNs, thus, if we can use GPU to accelerate arbitrary neural networks, we can have considerable gains in speed and both memory and power efficiency.

Graphics processing units were originally developed as a co processor alongside the main CPU, to offload the processing of graphics related tasks which are massively parallelizable. With co-accelerated advancements in games and GPU hardware, the GPU architecture reached a point where they could be used for general purpose computation. Initially, problems for general purpose computation in GPU (GPGPU) had to be converted into OpenGL shader programs utilizing non standard methods [10] but with eager support from GPU manufactures, platforms for general purpose computation on GPU were created. Compute Unified Device Architecture (CUDA) [11] and OpenCL [12] being two prominent examples. With vendor support and capable frameworks, GPUs have become a useful tool for massive data parallel problems.

This new found power of general purpose computation on GPUs have been extensively utilized in neural network development and research. One of the reason for deep learning growth has been attributed to advancement in GPU capability [13]. There are numerous GPU libraries supporting neural network acceleration [14] and frameworks around them [15]. But GPU acceleration on arbitrary structure neural networks is still lacking; which we explore in this paper. For GPU acceleration of ASNN, we first perform a pre processing step that segregates all the nodes into dependency hierarchy which we call layers. Then we use thread parallelization available with CUDA interface to compute all the nodes in that layer at the same time. With our methodology, we have shown that we can get significant speedup with the use of GPUs and the speedup gets better with respect to increase in neural network connections and depth of the network.

Remainder of this paper is organized as follows. Section 41.2 describes previous approaches related to our current work, Sect. 41.3 describes the neural network representation,

sequential and GPU activation methodology and experimental setup. In Sect. 41.4, we describe our results and compare the timing and speedup between two strategies. lastly in Sect. 41.5, we draw our conclusions and explain possible future directions.

41.2 Related Work

GPUs have been used for general purpose computation from before the time they provided formal support for it. Before APIs like CUDA and OPENCL were offered by the GPU vendors, researchers utilised OPENGL shader languages to coerce the GPU into performing non graphical processing [10, 16]. The application of GPUs for accelerating data parallel tasks has only increased after the introduction of the supported APIs. NVIDIA maintains a host of different libraries targeted to various application domains like deepLearning, signal processing, linear algebra and others [17].

The power of GPU have been eagerly utilized in the machine learning field as machine learning requires crunching through big numerical calculations and large number of iterations for making sense out of big datasets [18]. GPUs have been applied to Neural networks, and have been especially useful with the advent of deep learning, which uses neural networks with large number of neurons and hidden layers [19]. Previous work on implementation of neural networks in GPUs have started before the introduction of CUDA, where the authors utilized texture processing pipeline of the GPU to accelerate Multi layer perceptron and self organizing maps with significant speedup [20]. Scherer et. al. have shown in [21] that GPU can have gains of up to two orders of magnitude for convolution neural networks. Cheltur et al. have also shown that convolutional neural networks can be efficiently computed in GPUs with data framing in a form of matrix and performing matrix multiplication to compute the network [22]. Coates et al. have shown that many consumer grade GPUS in separate machines can be used for acceleration of convolutional neural networks by using CUDA, and using openMPI for multi GPU coordination [23]. Zhang et al. have also looked at accelerating sparse neural networks with custom hardware accelerators [24].

Other types of neural networks have also seen good results from GPU implementation. Nageswaran et al. have implemented a configurable simulation environment for the efficient simulation of large scale spiking neural networks on GPU [25]. Juang et al. have also shown significant speedup on fuzzy neural networks with high dimensional inputs by using parallel processing on GPUS [26], and GPUS have also been able to reduce recurrent neural networks training time by a factor of 32 [27].

Neural network in general form are also graph structures, and there have also been numerous research on graph pro-

cessing on GPUs. Luo et al. showed that the speedup of up to $10\times$ could be achieved with GPU implementation for breath first search [28]. Harish and Narayan give implementation of various graph processing algorithms on GPU in [29] and note that in some cases sequential approach does not transfer well to the GPU approach. In this paper, we are looking into networks with non uniform structure and have to perform a pre processing step on that structure to segregate the nodes where we have to apply graph processing approaches.

41.3 Methodology

41.3.1 CUDA

The CUDA application programming interface provides a way to structure our operations to run on Nvidia GPUs. The memory model in CUDA is divided into grids, blocks and threads which have access to specific kinds of memory and are all interfaced with the CPU, called a host, via a PCI bus as shown in Fig. 41.3. Code execution can be segmented to run in grids and blocks, both of which can be molded to have one to three dimensions depending on the problem. Each block runs the kernel, a block of CUDA procedure, in individual threads. Threads within a block can share a portion of memory called shared memory, which has very low latency as it resides on the chip. Current GPUs can run 32 threads in

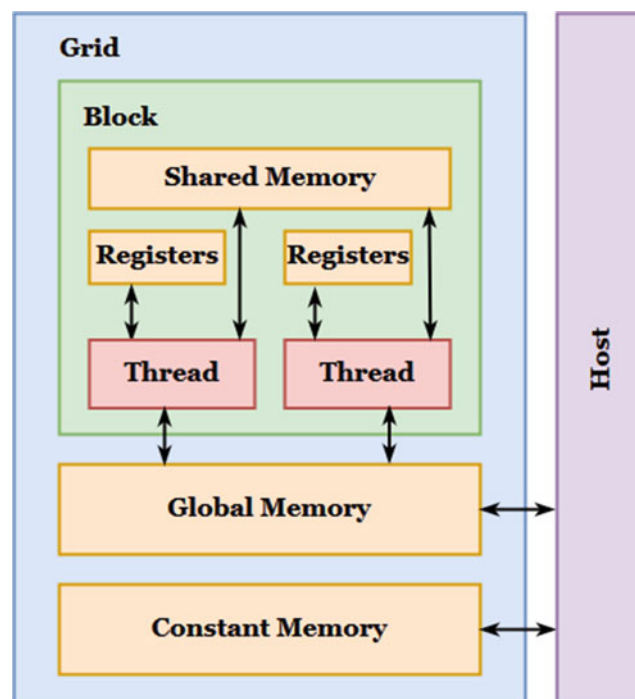


Fig. 41.3 Cuda memory architecture showing the relation between grid, block and threads and the corresponding proximity and connections of different kinds of memory elements

a block, which is also called a warp, at a single time where same instructions of a kernel executes on all the threads but runs on different data. Optimizing for efficient allocation of warps could lead to better performance. Threads in a block can be synchronized with `__syncthreads()` API call which syncs all the threads in a block at the code location where all of them execute `__syncthreads()`. Concept of unified memory was introduced in CUDA 4.0 which does away with the manual process of memory copying from device to host and back. Now, we allocate a portion of memory that is shared between both device and host and GPU driver takes care of transfer of data when data is accessed from either device. We use `cudaMallocManaged()` to allocate shared memory and use `cudaDeviceSynchronize()` before accessing any device data from the host.

41.3.2 Sequential Activation

Algorithm 1 Network segmentation algorithm

```

1: function SEGMENT_NETWORK(R, IN, OP, CON) ▷ R=required,
   IN=inputs, OP=outputs, CON=connections
2:   L ← []
3:   s ← IN
4:   while True do
5:     c = { b for (a,b) in CON if a in s and b not in s }
        ▷ Candidate nodes for the next layer
        ▷ Used nodes whose entire input set is in s
6:     t ← {}
7:     for n in c do
8:       all ← (for a in s for (a,b) in CON if b = n)
9:       if n in R and all then
10:        t ← t ∪ n
11:       end if
12:     end for
13:     if t = ∅ then
14:       break
15:     end if
16:     L ← L + t
17:     s ← s ∪ t
18:   end while
19:   return L
20: end function

```

For sequential activation of arbitrary neural network, we first perform pre processing on network structure to segment the network into sequential hierarchy of nodes, which we call layers of the ASNN. The algorithm to segment the network is given in Algorithm 1. The function takes all nodes, input nodes, output nodes and a structure containing all the network connections as input for processing. First, we find candidate nodes from the connections pool based on the nodes that have already been assigned to some layer. The candidates are those nodes for which incoming nodes all lie on the nodes that have already been assigned to a layer and all its outgoing nodes are

not in the assigned set. From the candidate set, we only add those nodes to the next layer if their entire input set is contained in the nodes which are already assigned a layer value.

For sequential propagation of the neural network, All the nodes starting of the input layer is sequentially activated till the output layer from which the answer is obtained. The activation is calculated by going through all the incoming nodes and multiplying the connection weights with the node values and then summing them and then squashing them with the sigmoid function.

41.3.3 Parallel GPU Activation

For single GPU activation of the neural network, we use the fact that all the nodes belonging to the same layer can be activated at once without compromising the output of the network in any way. For representation of the structure of network in the GPU, we use a custom data structure called `CudaNode` as depicted in Algorithm 2. Each `CudaNode` structure represents a single node of the neural network where each node contains a unique id, the number of incoming nodes, the integer array containing the node ids for the incoming connections and another float array for the corresponding weights for the incoming nodes. Then, we also have a Boolean that specifies if the node is a sensor i.e. takes external input for the network. The layer variable is set by the preprocessing of the neural network sequentially. The array of `CudaNode` structs are sorted in ascending order based on their layer number, where the input layer starts with value 0 and then climbs up to the last layer in the network. This is done to allow for better cache performance for unified memory in the GPU as input and output nodes will be close to each other in the array after being sorted.

Algorithm 2 Data structure to represent a single node in a network

```

1: struct CudaNode
2:   Integer: id           ▷ Unique id for the node
3:   Integer: layer       ▷ Layer the node is in
4:   Integer: numInNodes  ▷ No. of incoming nodes
5:   Boolean: isSensor    ▷ True if input node
6:   Integer[]: inNodes   ▷ Array of incoming Node ids
7:   Float[]: inWeights   ▷ Array of incoming Node weights
8: end struct

```

CUDA kernel for GPU activation is described in Algorithm 3. The kernel for GPU activation takes the value for total number of layers in the network, another integer array containing number of nodes in each layer, the main `CudaNodes` array containing sorted node entities, and an input array of floats which contains values for input layer of the neural network. Size of the output float array is equal to the size of number of nodes in the network as each node writes its

activation result to this array, which all other nodes will also be able to observe and write to. We have a variable cl which determines the current layer that the kernel is processing and sid , the start id which holds the id of the first node of the layer being computed. We will already have spawned many threads which will correspond to one node of the layer being activated. In case the current node is a sensor, we just perform sigmoid activation on the input variable from input array corresponding to the current node id. If the current node is not a sensor, we sum the values from all the incoming nodes after weighting them with the connection weight then do the sigmoid activation on the resulting sum.

Algorithm 3 CUDA kernel for calculating activation for ASNNs

```

1: ▷ Integer: TL = Total layers in a network
2: ▷ Integer[]: NNL = Number of nodes in layers
3: ▷ CudaNode[]: n = Array of all the nodes
4: ▷ Float[]: in = Input values for network
5: ▷ Float[]: op = Array for output values
6: function CUDA_ACTIVATION(TL, NNL, n, in, op)
7:   Integer: cl ← 0                                ▷ Current Layer
8:   Integer: sid ← 0                                ▷ Start id
9:   Integer: id ← threadIdx.x
10:  while cl < TL and id < NNLcl do
11:    CudaNode: cn ← nsid+id
12:    if cn is a sensor then
13:      open.id ← call sigmoid(inen.id)
14:    else
15:      Float: sum ← 0
16:      for i from 0 → cn.numInNodes - 1 do
17:        sum += cn.inWeightsi * open.inNodes[i]
18:      end for
19:      open.id ← call sigmoid(sum)
20:    end if
21:    call __syncthreads()
22:    sid ← sid + NNLcl
23:    cl ← cl + 1
24:  end while
25: end function

```

After one activation of one node is completed, we call `__syncthreads()` in the kernel to wait for all other nodes on the layer to finish computing. After synchronization, we increase the current layer variable to denote that we have progressed one layer of the neural network and increase start id by the number of nodes which were present in the completed layer. We compare current layer variable against total number of layers of the neural network and exit out of the loop if current layer is greater than the total layers, which signifies that the network has completed activation. After completing activation, we make sure to call `cudaDeviceSynchronize()` function before we read in the answers from the host to give the host enough time to copy the results from the device memory to host memory. Then, we read in the values from output array which has the final activations of output nodes in the network.

41.3.4 Experimental Setup

We used the CUBIX machine in our department for running all our experiments, which had the following configuration:

- Two 6 core Intel Xeon CPUs (E5-2620 0 @ 2.00GHz)
- CPU caches of L1: 32K, L2: 256K, L3: 15360K
- 64 GB of RAM
- 8 GTX 1080 with 8GB DRAM each

For all the results that follow, experiments were run 10 times and averaged for GPU activation timings and run 5 times and averaged for sequential activation timings.

41.4 Results and Discussion

41.4.1 Sequential Results

From Fig. 41.4 we can see that the execution of networks take linear amount of time with respect to the number of connections in the network. Increase in number of layers also correspond with the increase in execution time. Thus, with large number of connections and many layers, the execution time drastically increases. It is also possible for a network with same number of connection to have different execution time based on number of layers, as even with same number of connections, network with deeper layers take longer to execute.

41.4.2 Single GPU Result

From Fig. 41.5 we see that there is a general upward trend for the execution time with respect to the number of connections. But it should be noted that the slope is very flat compared to the sequential execution graph. The minimum is at 1 ms and the maximum at 2.5 ms.

41.4.3 Comparison and Speedup

From Fig. 41.4 we notice that compared to sequential execution, the GPU execution time lies flat and skims the x -axis. We can also compare the log of execution time from Fig. 41.6 where we see that the sequential execution time is very low for smaller networks but grows log linear with the number of connections. The log graph of execution time for sequential method has a steep slope at start, and still has positive slope after 30,000 connections, while the GPU execution curve is flat with the x axis after 30,000 connections. Thus, we can be certain that the GPU approach will scale well with increase in number of connections.

Fig. 41.4 Increase in execution time with respect to number of connections in a network for Sequential (blue) and GPU (red) approaches

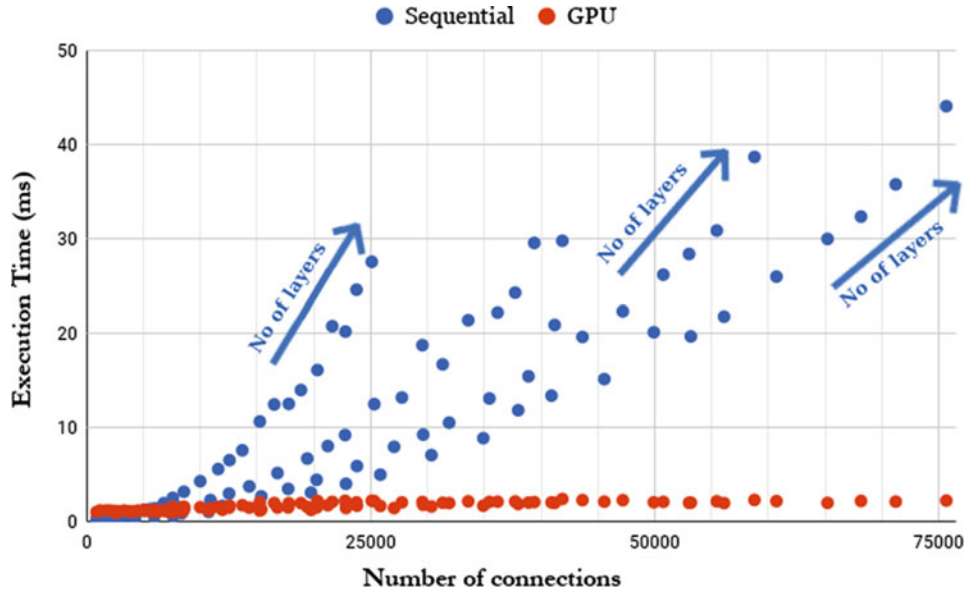


Fig. 41.5 Increase in execution time with respect to number of connections in a network using GPU implementation

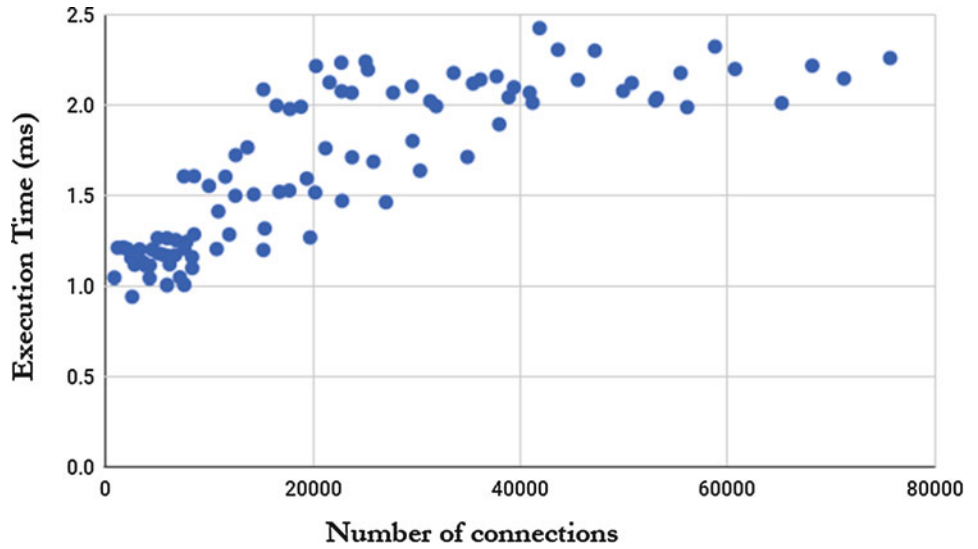


Fig. 41.6 Log of execution time with respect to number of connections in network, showing the comparison of execution times for Sequential and GPU approaches

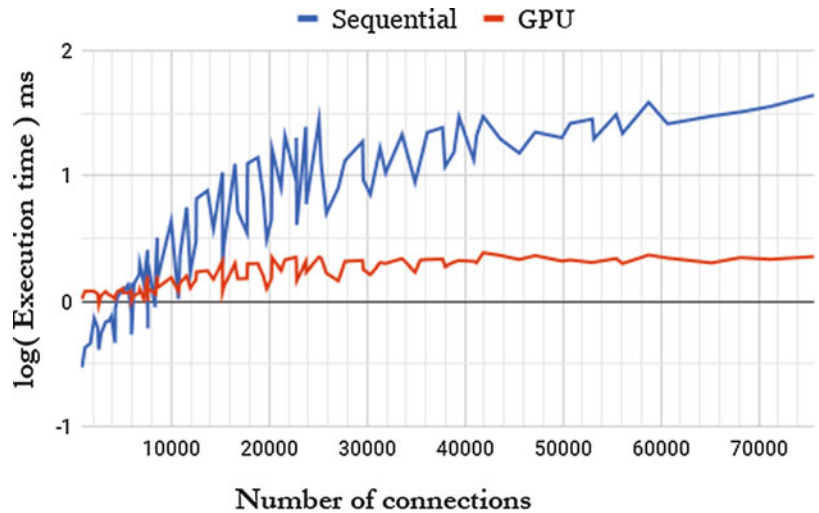
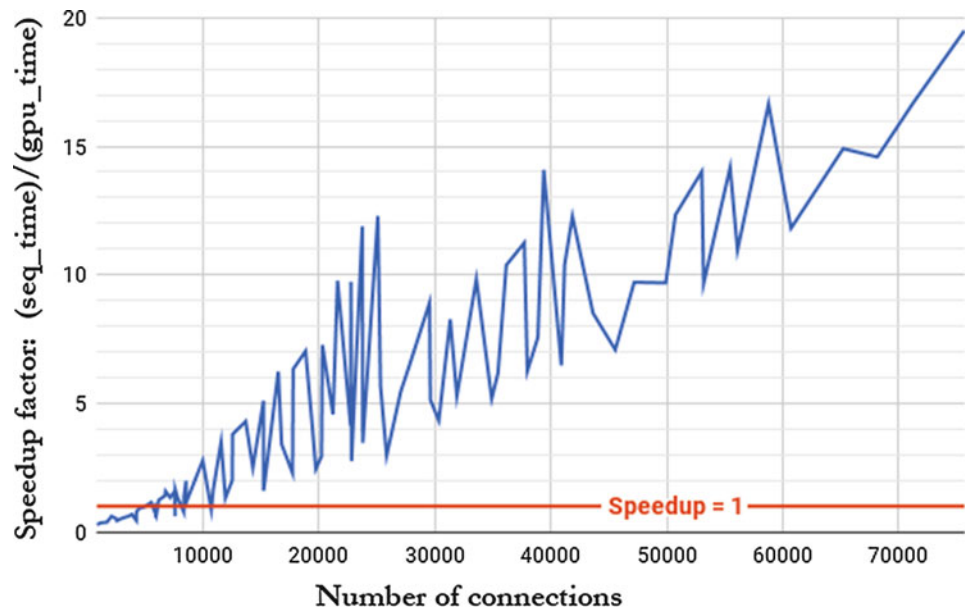


Fig. 41.7 Increase in speedup factor with increase in number of connections in a network showing the linear increase in speedup. Red line shows where speedup crosses the factor of 1



For speedup, we can see from Fig. 41.7 that initially, the speedup factor of sequential timings compared with GPU timings are lower than one which means that the GPU approach is slower than the sequential approach for very small networks. This is predictable as overhead of copying to and from the device negates any speedup from the computation happening at the device. Till 8000 connections, speedup from GPU is similar to the the sequential implementation, but after 8000 connections, speedup shows linear increase with number of connections. The jagged structure of the line is due to the variation in number of layers possible for a same connection count i.e. low depth networks can be computed quickly compared to deeper networks. From the graph, we can see that the speedup factor is fairly high for low depth networks too and is linearly increasing with increase in number of connections. This results signify that we will have more gains the larger our network gets. If we take a network with 70,000 connections, we can get up to 15 times speedup, which will have huge gains given that the networks are evaluated for thousands of iterations for any given problem.

41.5 Conclusion and Future Work

Our research focused on finding effective ways of accelerating arbitrary structured neural networks. We were able to show that: by pre-processing the network to segment it into dependent layers and then using CUDA threads to execute all the nodes in the same layer at the same time, we can get speedup that increases with the size of the connections in the neural network. From our experimentation, we have shown linear speedup increase for our GPU implementation compared to our sequential implementation.

We can further improve on this work by extending the approach to incorporate grid wide thread locking to synchronize threads in a grid group which will significantly increase the number of nodes which can be processed simultaneously. The natural extension of this work is, to find ways to perform network segmentation in GPU itself; which will also have significant impact on the overall efficiency of current approach. Our approach of using GPUs to accelerate arbitrary neural networks can also be used for other domains of research, as networks found in nature generally have non uniform structure, so our research can be incorporated for their study and simulation. One prominent example is of biological brains which have arbitrary structure with huge number of nodes and connections. We could simulate and study such large networks if we can extend the processing capacity by coordinating multiple GPUs to compute a single network.

Acknowledgments This material is based in part upon work supported by the National Science Foundation under grant number IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**(4), 115–133 (1943)
2. R. Hecht-Nielsen, Theory of the backpropagation neural network, in *Neural Networks for Perception* (Elsevier, Amsterdam, 1992), pp. 65–93
3. K. Fatahalian, J. Sugeran, P. Hanrahan, Understanding the efficiency of GPU algorithms for matrix-matrix multiplication, in *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware* (ACM, New York, 2004), pp. 133–137

4. Y. LeCun, J.S. Denker, S.A. Solla, Optimal brain damage, in *Advances in Neural Information Processing Systems* (Morgan Kaufmann, San Francisco, 1990), pp. 598–605
5. B. Hassibi, D.G. Stork, Second order derivatives for network pruning: optimal brain surgeon, in *Advances in Neural Information Processing Systems* (Kaufmann, San Mateo, 1993), pp. 164–171
6. K.O. Stanley, R. Miikkulainen, Evolving neural networks through augmenting topologies. *Evolut. Comput.* **10**(2), 99–127 (2002)
7. K.O. Stanley, D.B. D'Ambrosio, J. Gauci, A hypercube-based encoding for evolving large-scale neural networks. *Artif. Life* **15**(2), 185–212 (2009)
8. D. Floreano, P. Dürr, C. Mattiussi, Neuroevolution: from architectures to learning. *Evolut. Intell.* **1**(1), 47–62 (2008)
9. R. Miikkulainen, J. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, N. Duffy et al., Evolving deep neural networks (2017). Preprint. arXiv:1703.00548
10. F. Ino, J. Gomita, Y. Kawasaki, K. Hagihara, A gpgpu approach for accelerating 2-d/3-d rigid registration of medical images, in *International Symposium on Parallel and Distributed Processing and Applications* (Springer, New York, 2006), pp. 939–950
11. D. Kirk et al., NVIDIA CUDA software and GPU parallel computing architecture, in *ISMM*, vol. 7 (2007), pp. 103–104
12. J.E. Stone, D. Gohara, G. Shi, OpenCL: a parallel programming standard for heterogeneous computing systems. *Comput. Sci. Eng.* **12**(3), 66–73 (2010)
13. D. Strigl, K. Kofler, S. Podlipnig, Performance and scalability of GPU-based convolutional neural networks, in *2010 18th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (IEEE, New York, 2010), pp. 317–324
14. C. Nvidia, *Cublas Library*, vol. 15(27) (NVIDIA Corporation, Santa Clara, CA, 2008), p. 31
15. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., Tensorflow: a system for large-scale machine learning, in *OSDI*, vol. 16 (2016), pp. 265–283
16. D. Göttsche, GPGPU-basic math tutorial. Univ. Dortmund, Fachbereich Mathematik, 2005
17. NVIDIA, GPU-accelerated libraries for computing (2018) [Online]. Available: <https://developer.nvidia.com/gpu-accelerated-libraries>
18. D. Steinkraus, I. Buck, P. Simard, Using GPUs for machine learning algorithms,” in *Proceedings of Eighth International Conference on Document Analysis and Recognition, 2005* (IEEE, New York, 2005), pp. 1115–1120
19. I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, *Deep Learning*, vol. 1 (MIT Press, Cambridge, 2016)
20. Z. Luo, H. Liu, X. Wu, Artificial neural network computation on graphic process unit, in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. IJCNN'05*, vol. 1 (IEEE, New York, 2005), pp. 622–626
21. D. Scherer, H. Schulz, S. Behnke, Accelerating large-scale convolutional neural networks with parallel graphics multiprocessors, in *International Conference on Artificial Neural Networks* (Springer, New York, 2010), pp. 82–91
22. S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, E. Shelhamer, cuDNN: efficient primitives for deep learning (2014). Preprint. arXiv:1410.0759
23. A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, N. Andrew, Deep learning with COTS HPC systems, in *International Conference on Machine Learning* (2013), pp. 1337–1345
24. S. Zhang, Z. Du, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, Y. Chen, Cambricon-X: an accelerator for sparse neural networks, in *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* (IEEE, New York, 2016), pp. 1–12
25. J.M. Nageswaran, N. Dutt, J.L. Krichmar, A. Nicolau, A.V. Veidenbaum, A configurable simulation environment for the efficient simulation of large-scale spiking neural networks on graphics processors. *Neur. Netw.* **22**(5–6), 791–800 (2009)
26. C.-F. Juang, T.-C. Chen, W.-Y. Cheng, Speedup of implementing fuzzy neural networks with highdimensional inputs through parallel processing on graphic processing units. *IEEE Trans. Fuzzy Syst.* **19**(4), 717–728 (2011)
27. X. Chen, Y. Wang, X. Liu, M.J. Gales, P.C. Woodland, Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch, in *Fifteenth Annual Conference of the International Speech Communication Association* (2014)
28. L. Luo, M. Wong, W.-M. Hwu, An effective GPU implementation of breadth-first search, in *Proceedings of the 47th Design Automation Conference (ACM, New York, 2010)*, pp. 52–55
29. P. Harish, P. Narayanan, Accelerating large graph algorithms on the GPU using CUDA, in *International Conference on High-Performance Computing* (Springer, New York, 2007), pp. 197–208

Parallelizing the Slant Stack Transform with CUDA

42

Dustin Barnes, Andrew McIntyre, Sui Cheung, John Louie, Emily Hand, and Frederick C. Harris Jr.

Abstract

In geophysics, the slant stack transform is a method used to align signals from different sensors. We focus on the use of the transform within passive refraction microtremor (ReMi) surveys, in order to produce high resolution slowness-frequency plots for use as samples in a machine learning model. Running on a single central processing unit (CPU) thread, this process takes approximately 45 min, 99.5% of which consists of the slant stack transform. In order to reduce the time taken to perform the transform, we use NVIDIA CUDA programming model. Using the same CPU, augmented with a GeForce RTX 2080 Ti we were able to reduce this time down to as little as 0.5 s.

Keywords

Parallel programming · GPU · CUDA · PyCUDA · Refraction microtremor (ReMi) · Seismic refraction · Slant stack · Radon transform · Beamforming transform · Rayleigh waves

42.1 Introduction

Refraction microtremor (ReMi) surveys use geophone-array recordings of the Rayleigh component of ambient, vertically directed seismic ground-vibration noise from passive

D. Barnes · A. McIntyre · S. Cheung · E. Hand · F. C. Harris Jr. (✉)
Computer Science and Engineering, University of Nevada, Reno,
Reno, NV, USA
e-mail: dkbarnes@nevada.unr.edu; amcintyre@nevada.unr.edu;
scheung@nevada.unr.edu; emhand@unr.edu; fred.harris@cse.unr.edu

J. Louie
Nevada Seismological Laboratory, University of Nevada, Reno, Reno,
NV, USA
e-mail: louie@seismo.unr.edu

sources in order to obtain a shear-wave profile, describing the seismic-velocity property of soils and rocks at different depths [1]. ReMi does not require drilling or the use of a seismic source—such as hammers or explosives. ReMi provides a cheap, easy to set up survey method that is effective in urban environments. Although ReMi was originally developed as a means of assessing earthquake safety and construction code compliance for building sites, ReMi has been used for a wide variety of applications such as geologic basin and bedrock analysis, and foundation design for sites such as bridges and wind turbines.

Conducting a ReMi survey require obtaining array recordings, transforming the time- and distance-dependent seismic records into a slowness-frequency (p-f) plot, manually estimating the fundamental-mode Rayleigh dispersion curve, and then manually fitting the selected dispersion points to a 1D shear-velocity versus depth model. This is a labor-intensive process with a mathematically under-determined result, which can vary depending on the individual performing the analysis. In addition, the amount of raw seismic data being recorded for analysis is constantly increasing, beyond the capacity of human analysts to keep up. As such, the ultimate goal of this research is to generate a machine learning model capable of generating a velocity model given a p-f plot.

Methods of generating p-f plots are designed for human consumption, constrained by the limitations of decades old hardware. Increasing the demands on this algorithm to create samples suitable for consumption by a machine learning model make the time required to generate a batch of samples infeasible for training. The vast majority of this time is the result of performing the slant stack transform, which aligns the signals from the geophone sensors along the time-offset (τ) domain. Figure 42.1 is an example of visualized waves in geophysics. It is built from the sensor measurements in a linear array (traces). It accumulate samples (intensity) at each

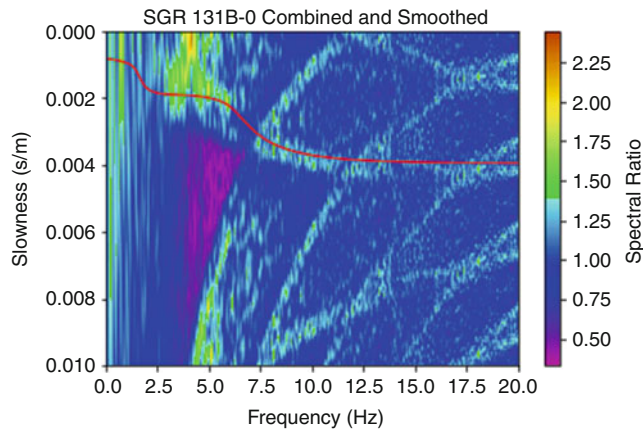


Fig. 42.1 An example of a generated p-f plot, with the fundamental mode Rayleigh dispersion curve drawn in red. This one was generated using the serial implementation

point and it is currently in a low resolution grids for program performance purpose.

The remainder of the paper is structured as follows: Sect. 42.2 provides an overview of the ReMi process and seismic imaging techniques. Section 42.3 provides a brief background for CUDA. Section 42.4 provides a description of the algorithm which slant stacking utilizes, and different methods of optimizing this for CUDA. Section 42.5 provides details on hardware and software used. Section 42.6 describes the outcomes of our study, and Sect. 42.7 provides conclusion and a discussion on future work, potential improvements, and applications of our work.

42.2 Background

Several other works have shown the potential for CUDA accelerated computation within geophysics. Refs. [2] and [3] are both examples of using graphics processing units (GPU) to accelerate processing, while [4] shows how ray tracing on modern hardware can be used to increase the fidelity of computational models. No work has been done on accelerating the slant stack transform or ReMi surveys, with most research in the field instead investigating different areas in which the technique may be applied.

The remainder of this paper focuses on generating p-f plots and the slant stack transform. Generating a p-f plot consists of several distinct steps. First, the raw signals must be processed. Each trace, consisting of the raw signals from each geophone, must be centered, ensuring that they are measured as variations around a zero-volt response level, and RMS normalization is performed. Next, the signal is transformed into the frequency domain and filtered. At minimum, a bandpass filter is used with the maximum frequency determined by the Nyquist frequency—the highest frequency able to be

reliably sampled, based on the sampling rate of the sensors—using a bandpass filter. After frequency domain filtering is performed, the signal for each trace is then transformed back into the time domain.

Once the raw data has been preprocessed, the slant stack transform is applied. This generates a p- τ plot, with axes of p, or slowness (s/m) and τ , or zero offset reflection time, and with each point on the grid representing the intensity of the signal received.

$$\tau = t - p * d \quad (42.1)$$

$$t = \tau + p * d \quad (42.2)$$

As shown in Eq. (42.1), τ is calculated by taking the distance (d) of the current sensor from the first sensor in a linear array and multiplying it by the current slowness (p) to align the signal from a distant geophone to that of the first. The intensity of the signal is the added to the current grid point. When generating the p- τ plot, every point in the plot is iterated over, and the time in the trace is calculated using Eq. (42.2). As the calculated trace time is unlikely to align exactly with a sampled point on the trace, linear interpolation is performed between the two nearest values.

In the case where a time is outside the bounds of the trace, either with negative time or beyond the sampled time frame, no intensity is added. In addition to creating a plot from slowness $p = 0$ to p_{max} , a p- τ plot is also created ranging from $-p_{max}$ to $p = 0$. Thus, the plot ranging from $[0, p_{max}]$ captures waves travelling along the array from the first sensor to last, and the plot ranging from $[-p_{max}, 0]$ captures waves travelling along the array from the last sensor to the first. These are referred to as the forward and reverse plots respectively.

Once the forward and reverse p- τ plots have been calculated, the reverse plot is inverted along the slowness axis, and added to the forward plot, creating the combined plot. A final Fourier transform is applied to all three plots, transforming them into the frequency domain along the x axis. Additional work may be done in order to visualize the data in different ways.

42.3 CUDA

The graphical processing unit (GPU) has become a popular device for creating a high performance parallel computing platform for a relatively low cost. Each GPU contains a large number of processing cores and each core can create many threads that can all be executed in parallel. As a result, many different types applications utilize GPUs to solve computationally expensive problems and have been successful in increasing the performance. NVIDIA provides their Compute Unified Device Architecture (CUDA) programming model in order for new and existing applications to be executed on the

GPU. CUDA extends the standard C/C++ language to give direct access to the instructions and memory management for parallel computation on NVIDIA GPUs

Starting with the G80 series, NVidia unveiled a new architecture called the Compute Unified Device Architecture (CUDA) to ease the struggles of programmers attempting to harness the GPU's computing power [5]. While the new architect did not change the pipeline for graphics programmers, it did unify the processing architecture underlying the whole pipeline. Vertex and fragment processors were replaced with groups of Thread Processors (CUDA Cores) called Streaming Multiprocessors (SM). Initially with the G80 architecture, there were 128 cores grouped into 16 SMs. The Kepler architecture has 2880 cores grouped into 15 SMs. The Maxwell has SMs with 128 cores and has a varying number of SMs (3–15) depending upon the model of GPU. The Pascal architecture was also 128, but Turing was 64, and Ampere has gone back to 128

In addition to the new architecture, CUDA also included a new programming model which allowed application programmers to harness the data parallelism present on the GPU. The primary abstractions of the programming model are kernels and threads. Each kernel is executed by many threads in parallel; CUDA threads are very lightweight and allow many thousands of threads to be executing on a device at any given time.

CUDA exposes the computational power of the GPU through a C programming model. It additionally provides an API for scheduling multiple streams of execution on the GPU. This allows the hardware scheduler present on CUDA enabled GPUs to more efficiently use all of the compute resources available. When using streams, the scheduler is able to concurrently execute independent tasks.

Changes over the years to CUDA have included shared memory between the host and device as well as kernel calls from kernels.

42.4 Approach/Implementation

The slant stack algorithm itself is fairly straightforward. Algorithm 1 describes the basic serial implementation of the transform, with Algorithm 4 describing the processes used to interpolate the intensity of a signal on a trace, given the ideal time. This process is identical for both the serial and parallel algorithms, however it does require a significant amount of computation. Our test case, generating a 1648×1648 plot from a stream with 15 traces, requires 147 million executions of Algorithm 4.

For our parallel implementation, two algorithms are described. Algorithm 2 performs a CUDA call on each trace in a stream, limiting the amount of memory required at the cost of more frequent CUDA calls. Algorithm 3 sends the full

Algorithm 1 Serial slant stack transform

```

Create Output Plot
for trace do
  for tau do
    for p do
       $t_{ideal} = \tau + d * p$ 
      Accumulate on plot // Call Alg 4
    end
  end
end
end

```

stream to the GPU, and processes it in a single CUDA call, but there must be sufficient VRAM available on the device for the output array, all traces, as well as memory available for computation.

When Algorithm 2 is executed, a trace is sent to the GPU and the result of each point in the output array is calculated and summed to the same global array. Once all traces have been called, the results are transferred back to the host. This method simplifies the process—each trace modifies each point of the output array once, and so it is guaranteed that there will be no race conditions present when calculating a single trace—as well as reducing the amount of memory required for computation. For our test case, the output array requires approximately 1.28 GB of VRAM to store, with additional memory needed for the trace itself. Some additional cost is also incurred due to the more frequent communication required between the device and host, though it is not significant unless generating very small plots.

Algorithm 2 CUDA Kernel per Trace

```

CPU:
Create Output Plot
for trace do
  CUDA Call
end
end

KERNEL:
 $t_{ideal} = \text{threadIdx.x} + \text{threadIdx.y} * \text{distance}$ 
Add to Output Array // Call Alg 4

```

Algorithm 3 performs a batch computation of all traces. Similar to Algorithm 2, a global output array is used to store the results. However consideration must be given to potential data races, as multiple threads may attempt to modify the same memory locations. In this case, each trace within the stream will modify each element of the global output array once—resulting in 15 increments of every value. This requires additional checks to ensure that no two threads are operating on the same location, either via use of atomics, temporary arrays for each trace, or subdividing the output array into several smaller arrays. Additionally, loading the entirety of the stream into VRAM may cause difficulties

Algorithm 3 CUDA kernel per stream

```

CPU:
Create Output Plot

Send all traces to GPU
CUDA Call

KERNEL:
  Calculate Trace Time

   $t_{ideal} = \text{threadIdx}.x$ 
  +  $\text{threadIdx}.y * \text{distance}$ 
  +  $\text{threadIdx}.z * \text{traceLen}$ 
  Add to Output Array // Call Alg 4

```

Algorithm 4 Process to find intensity

```

if  $abs(t_{ideal}) < t_{max}$  then
   $t_{low} = \text{floor}(t_{ideal}/dt)$ 
   $t_{high} = \text{ceil}(t_{ideal}/dt)$ 
   $\text{intensity} = \frac{\text{trace}[t_{high}] - \text{trace}[t_{low}]}{t_{high} - t_{low}} (t_{ideal} - t_{low})$ 
   $\text{plot}[\tau][p] += \text{intensity}$ 
end
else
  Skip
end

```

depending on the size of the survey. For the purposes of this study, we only implement Algorithm 2.

To implement these methods, we use the PyCUDA library, which gives access to NVIDIA's CUDA parallel computation API in python. This allows us to easily parallelize our serial code, which is implemented in python. Additionally, utilizing python allows for easier integration with machine learning models, and better support for reading geophysics data.

42.5 Hardware and Software Used

CUDA and PyCUDA are used in this project both in the sequential and parallel implementations. As mentioned in Sect. 42.3, CUDA is parallel computing platform and programming model that was designed for NVIDIA graphics cards. CUDA helps speed up computing application significantly when there is a lot of data that is computed at the same time. CUDA supports languages such as C, C++, Fortran, Python, and Matlab. In this work, we utilized PyCUDA for a python implementation of the slant stack transform implementation [6].

PyCUDA is a Pythonic access method for NVIDIA's CUDA parallel computation API. PyCUDA has C++ style syntax so that it's ease-to-use. It provides automatic error checking where all CUDA calls are translated as Python exceptions. It has a very fast runtime, along with automatically allocating and de-allocating space in the program [7].

42.5.1 Hardware Used

The hardware described was used for both the sequential and parallel implementations of the code. Since CUDA is limited to use for NVIDIA graphics cards. It is very important to have a workstation that is set up with NVIDIA graphics card(s). We also wanted to have a decent CPU where it can send the CUDA instructions to the NVIDIA graphics card(s) rapidly. For this paper, we used a workstation that has NVIDIA graphics card and a Intel CPU. Here is the essential components of our workstation set up:

- Intel® Core™ i5-4670K Processor
- GeForce RTX 2080 Ti

Intel® Core™ i5-4670K Processor is the 4th generation Intel® Core™ i5 Processors. It has 4 cores and 4 threads, processor base frequency of 3.40 GHz, and 6 MB of Intel® Smart Cache (which allows all cores to dynamically share access to the last level cache) [8].

The machine was running a Debian version of Linux.

GeForce RTX 2080 Ti was the latest version of NVIDIA graphics card at the time of this work. A GeForce RTX 2080 Ti was used in the workstation and it is made by EVGA and the model is 11G-P4-2281-KR. A GeForce RTX 2080 Ti has 4352 of NVIDIA CUDA® Cores, 14 Gbps of memory speed, 11 GB GDDR6 VRAM.

42.5.2 Software Used

In the CUDA implementation, we used software packages from *CUDA 10.2*, *pycuda 2019.1.2*[7], *NVIDIA Driver 3.5*, *numpy 1.18.2*[9], *obspy 1.2.1*[10,11], and *evodcinv 1.0.0*[12]. ObsPy is a seismology framework for Python, which provides tools necessary to parse common file formats. In this project, ObsPy is exclusively used to read in the trace data.

42.6 Results

Implementing the slant stack transform in CUDA allowed for a considerable speed up, while still producing accurate results. Prior to CUDA parallelization, an individual trace would take approximately 400 s to calculate using the CPU described in Sect. 42.5. This resulted in our test case needing approximately one and a half hours of calculation. After parallelization was implemented, a single trace could be processed in approximately 0.01 s, with the entire process averaging 0.65 s. We observed an average speedup of approximately 10,000 times. This is consistent with the peak performance of the two devices, as the the 2080 Ti is capable

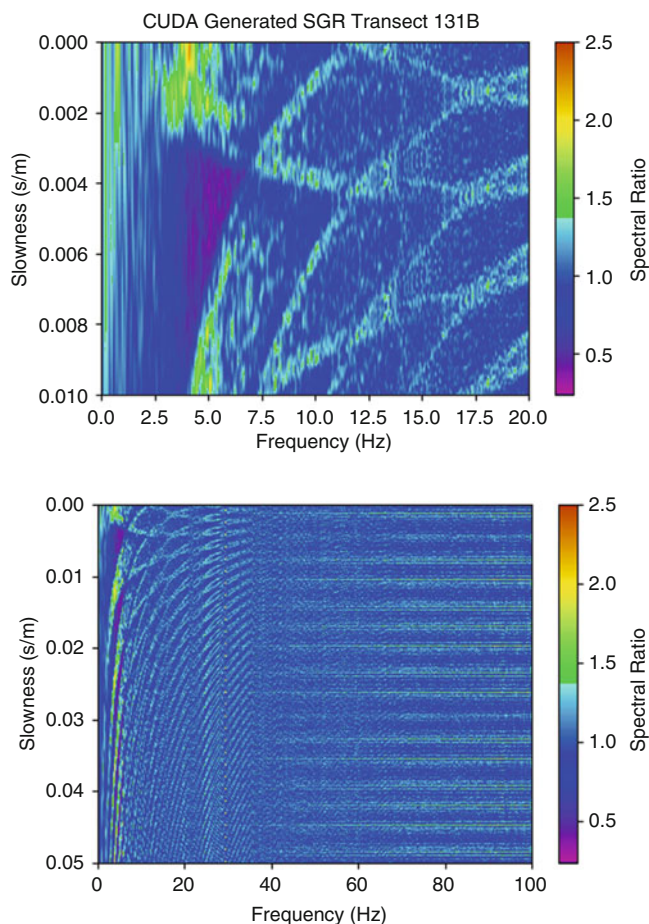


Fig. 42.2 These p - τ plots were generated using our CUDA implementation. The top plot shows the zoomed in figure, comparable to the one presented in Fig. 42.1. The bottom plot shows the full figure, extending to the Nyquist frequency of our test case

of 14.2 teraflops (TFLOPS), compared to the 13.6 gigaflops (GFLOPS) available when using a single core of on the CPU.

We measured both GPU memory and processor utilization using the NVIDIA System Management Interface. As discussed in Sect. 42.4, Algorithm 2 was unable to saturate the GPU, peaking at only 5% GPU utilization. Despite this, memory utilization peaked at 40%, or 4.4 GB. This shows that there is still room for improvement, such as processing multiple traces at once.

Figure 42.2 shows the CUDA generated plots, with one zoomed in, covering the same region as the serial version shown in Fig. 42.1. The second plot shows the same data, beyond the fundamental mode dispersion curve and extending to the Nyquist frequency of our test case. It's worth noting that these plots are both generated with the same resolution, however the larger one is impractical for use without using GPU acceleration.

Figure 42.3 shows the runtime for both devices at different problem sizes. The runtime scales linearly with the number of

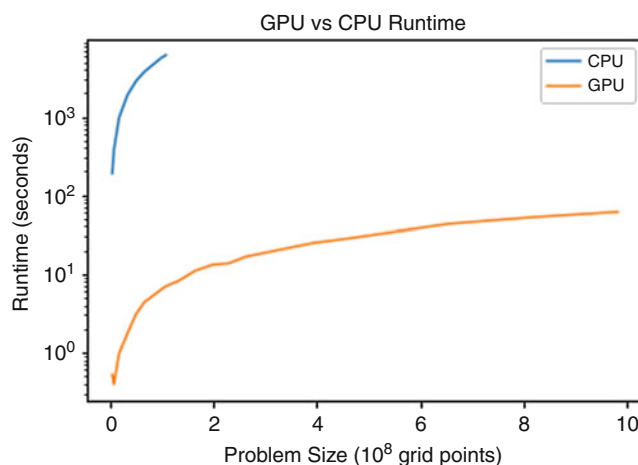


Fig. 42.3 This is the runtime comparison plot between the Intel® Core™ i5-4670K (CPU) and the GeForce RTX 2080 Ti (GPU)

grid points that need to be calculated, with some inconsistencies on the GPU when calculating with very small problem sizes.

42.7 Conclusion and Future Work

The findings in this project indicate that the spreading the calculations for the slant stack transform across multiple CUDA cores provides substantial speedup. The drastic increase in the amount of data that the program can produce in a certain amount of time makes the method viable for use in machine learning. By making the slant stack transform a parallel process through CUDA, we were able to achieve close to real time computation.

Although in this work we apply the slant stack transform to ReMi surveys, the transform is not specific to this technique. As such, this can be beneficial to any work which uses slant stacking, particularly if processing a large amount of files, or attempting to generate plots with high resolution.

Future work for this project includes having the calculations be performed through batch processing. This would significantly boost the output of the program, since more data can be generated over a shorter period of time. In addition to the slant stack transform, other parts of the algorithm can be parallelized. For example, the process of transforming and filtering the raw data is currently a serial operation. While the overhead from preprocessing everything in a serial fashion is small, larger inputs may negatively affect the runtime of the program. The Fourier transforms performed on the set of traces can also be implemented into CUDA as well, limiting the data that needs to be sent to the GPU to only the raw traces.

References

1. J. Louie, Faster, better: Shear-wave velocity to 100 meters depth from refraction microtremor arrays. *Bull. Seismol. Soc. Am.* **91**, 04 (2001)
2. Y. Wang, H. Zhou, X. Zhao, Q. Zhang, P. Zhao, X. Yu, Y. Chen, CuQ-RTM: a CUDA-based code package for stable and efficient q-compensated reverse time migration. *Geophysics* **84**(1), F1–F15 (2019) [Online]. Available: <https://doi.org/10.1190/geo2017-0624.1>
3. B. Holt, D. Ernst, Accelerating geophysics simulation using CUDA. *J. Comput. Sci. Educ.* **2**, 12 (2011)
4. M. Sarajaervi, H. Keers, Ray-based modeling and imaging in viscoelastic media using graphics processing units. *Geophysics* **84**(5), S425–S436 (2019) [Online]. Available: <https://doi.org/10.1190/geo2018-0510.1>
5. D. Kirk, NVIDIA CUDA software and GPU parallel computing architecture, in *ISMM*, vol. 7 (2007), pp. 103–104
6. J. Nickolls, I. Buck, M. Garland, K. Skadron, Scalable parallel programming with CUDA. *Queue* **6**(2), 40–53 (2008) [Online]. Available: <https://doi.org/10.1145/1365490.1365500>
7. A. Klöckner, N. Pinto, B. Catanzaro, Y. Lee, P. Ivanov, A. Fasih, GPU scripting and code generation with PyCUDA, in *GPU Computing Gems Jade Edition*, ed. by W. Mei, W. Hwu (Elsevier Inc., New York, 2012), pp. 373–385
8. Intel Corp., Intel® core™ i5-4670k processor: 6m cache, up to 3.80 GHz, visited on (11/2/2020) [Online]. Available: <https://ark.intel.com/content/www/us/en/ark/products/75048/intel-core-i5-4670k-processor-6m-cache-up-to-3-80-ghz.html>
9. NumPy, NumPy: The fundamental package for scientific computing with python, visited on (11/2/2020) [Online]. Available: <https://numpy.org/>
10. The ObsPy Development Team, ObsPy: A python framework for seismology, visited on (11/2/2020) [Online]. Available: <https://github.com/obsproxy/obsproxy/wiki>
11. M. Beyreuther, R. Barsch, L. Krischer, T. Megies, Y. Behr, J. Wassermann, ObsPy: a python toolbox for seismology. *Seismol. Res. Lett.* **81**(3), 530–533 (2010) [Online]. Available: <https://doi.org/10.1785/gssrl.81.3.530>
12. K. Luu, EvoDCinv, visited on (11/2/2020) [Online]. Available: <https://github.com/keurfonluu/EvoDCinv>

Part VIII

Social Computing/E-Learning

Recommender Systems Evaluator: A Framework for Evaluating the Performance of Recommender Systems

43

Paulo V. G. dos Santos, Bruno Tardiole Kuehne, Bruno G. Batista, Dionisio M. Leite, Maycon L. M. Peixoto, Edmilson Marmo Moreira, and Stephan Reiff-Marganiec

Abstract

Recommender systems are filters that suggest products of interest to customers, which may positively impact sales. Nowadays, there is a multitude of algorithms for recommender systems, and their performance varies widely. So it is crucial to choose the most suitable option given a situation, but it is not a trivial task. In this context, we propose the *Recommender Systems Evaluator (RSE)*: a framework aimed to accomplish an offline performance evaluation of recommender systems. We argue that the usage of a proper methodology is crucial when evaluating the available options. However, it is frequently overlooked, leading to inconsistent results. To help appraisers draw reliable conclusions, RSE is based on statistical concepts and displays results intuitively. A comparative study of classical recommendation algorithms is presented as an evaluation, highlighting RSE's critical features.

Keywords

Big data · Collaborative filtering · Evaluation metrics · Graphical analysis · Parameter optimization ·

Performance evaluation · Ratings · Recommender systems · Statistics · Workflow

43.1 Introduction

When asked what a recommender system is, most people would probably say that they do not know. Yet, it is very likely that they have already had some experiences with recommendations. Indeed, recommenders are pervasive, and can be found in e-commerce stores, streaming platforms, social networks, and others. There are also applications in other less related fields, such as suggesting training for software developers [1].

Recommender systems can have multiple forms, but these are generally arising from traditional approaches: content-based, collaborative filtering, or knowledge-based. The first one builds user/item profiles using textual information, like tags or product descriptions. The collaborative filtering approach relies on ratings to infer similarities between a pair of users/items. The last kind of approach recommends after successive interactions between the user and the platform. It uses knowledge retrieved from a specific domain [2, 3].

As the available algorithms perform differently in diverse scenarios, it is essential to choose the best option for any given situation due to the possible financial impact. The authors of [4] state that the rapid growth of information and clients poses three critical challenges for recommender systems: high-quality recommendations, many recommendations per second in big data scenarios, and high coverage even when facing data sparsity. The algorithm will have to adjust its parameters to extract the best possible results in a specific situation. Therefore, it is crucial to have ways to measure the performance of the recommender systems. Indeed, this kind of evaluation is one of the significant challenges in the field,

P. V. G. dos Santos (✉) · B. T. Kuehne · B. G. Batista · E. M. Moreira
Federal University of Itajubá, Itajubá, MG, Brazil

D. M. Leite
Federal University of Mato Grosso do Sul (UFMS), Ponta Porã, MS, Brazil

M. L. M. Peixoto
Federal University of Bahia (UFBA), Salvador, BA, Brazil
e-mail: maycon.leone@ufba.br

University of Campinas, Campinas, SP, Brazil
e-mail: maycon.leone@ufba.br

S. Reiff-Marganiec
University of Derby, Derby, UK
e-mail: S.Reiff-Marganiec@derby.ac.uk

and even though there are many metrics, how to choose the most suitable options given a situation is still unknown [5].

In this paper, we present the *Recommender Systems Evaluator (RSE)*, which executes an offline performance evaluation of the recommenders. We build the tool around a robust methodology for assessment to ensure that trusted and reproducible results are obtained. RSE has a well-organised set of metrics, and its outputs are presented intuitively. It can be used to either finding the optimal parameter values for a specific approach or for comparing and choosing among various existing implementations. Thus, we contribute to provide robust comparisons and reliable results. To validate the framework, a comparative study of traditional recommenders is shown. RSE is available for community research.¹

This paper is organized as follows: Sect. 43.2 shows related work found in the literature. RSE is explained in Sect. 43.3 and a comparative study is presented in Sect. 43.4. Finally, we draw conclusions in Sect. 43.5.

43.2 Related Works

We will do a brief description of some related works. *TagRec* [6] is an evaluation and development recommender framework with a wide set of ready to use algorithms. Its focus is on tag recommenders. *RiVal* [7] is a toolkit that allows transparent and complete control of the evaluation setting. The third software is also one of the most complete in the field. It is called *Lenskit* [8], and its focus is on collaborative filtering methods. The work of [9] uses plugins to ease the incorporation of new algorithms, and also has a user interface. However, he has a tight set of metrics.

All of the introduced approaches are valuable in evaluating recommender systems. Some of them use data loaded directly from files. Others produce results only in numerical form, causing more effort to interpret them.

To the best of our knowledge, the main differences of RSE when comparing to the related works are: (i) uses PostgreSQL and has a standardized structure to store test data; (ii) results are presented using mean and variability, so they are interpreted in a statistical point of view; (iii) aside from comparing algorithms, RSE provides ways to find the optimal parameter values for a single algorithm; (iv) has built-in graphical tools; (v) it does not focus on a specific group (like TagRec and Lenskit). This approach also has drawbacks but tends to serve broader situations.

43.3 RSE

RSE is a software written in Java that executes the offline performance evaluation of recommender systems. It enables us to have insights into performance to adjust recommender algorithm parameters and get the best possible operation. RSE can also perform comparisons between algorithms to determine which method achieves the best outcome in a given situation. Results are presented in an intuitive graphical way to be easily understood.

The software has two modes of user request distribution, which simulate typical situations and peak times, respectively. It was developed in a modular fashion aiming to be easily extendable. If the appraiser needs new features, small parts inside a module can be changed, leaving the other untouched.

Figure 43.1 presents the workflow of RSE. The operation flow follows the steps: First step is data preparation. Test data are imported into the standard structure of RSE. Then several values are calculated, like the overall mean and similarities, Second step involves choosing an algorithm. Users can pick between the existing ones or add a customized algorithm in the framework. Third step is the performance evaluation. The last step, the results are stored in the database, and then graphics and complementary information are generated.

43.3.1 DataBase

The purpose of this module is to encapsulate database access and separate it from other parts. RSE assumes that the simplest recommendation model needs users, items, and their relations (ratings). Other structures can be used depending on the recommendation method, like tags, similarity, and content vectors. It also has auxiliary tools to facilitate the import of data.

43.3.2 Recommender

In this module, there are four classic algorithms available to be used. Two used algorithms are variations of collaborative

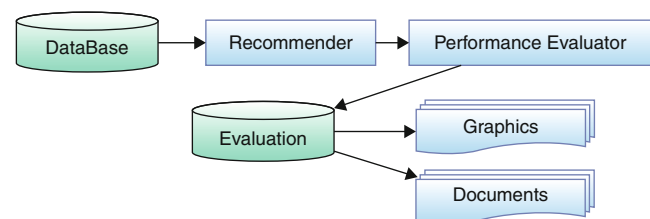


Fig. 43.1 Workflow of RSE

¹<https://github.com/paulovgs/RSEvaluator>

filtering (user-user and item-item), the third one is a content-based recommender, and the last one is hybrid. The classic approaches are still relevant in nearly all kinds of applications [2]. They serve as base for more complex algorithms, such as the ones used by Netflix. Also, RSE allows the appraiser to add new methods if wanted. Some plugins may be available in the future to import algorithms directly from the well-known recommendation frameworks.

43.3.3 Performance Evaluator

The main task of this module is to execute the evaluation itself. To optimize a running program, the JVM has some singularities. The most often used classes are cached, and the other ones may be loaded just in the moment of their need. Thus, the performance of a running program can vary sharply in the first instants of its execution. Therefore RSE has a warm-up method, where we do the same steps of a usual experiment, but the workload is reduced, and the results are dropped out. It is intended to stabilize the JVM and create a favorable environment for evaluation.

Due to the stochastic process of measuring, it is crucial to repeat the experiment a number of times. Otherwise incorrect conclusions could be taken. RSE provides methods to calculate the number of replicas required on each test, or it can be alternatively defined by the appraiser.

A configuration file is provided to set up tests, where the appraiser needs to make some choices like state the response variables and factors to be used. The setup file will guide the overall evaluation workflow, which is presented in Fig. 43.2.

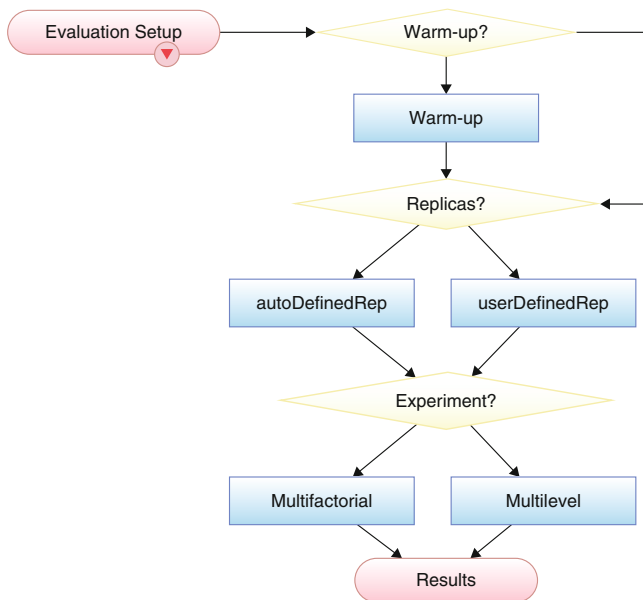


Fig. 43.2 Evaluation workflow

The reader is referred to [10] to get a deep understanding of performance evaluation, and the following sections will explore the available options in more detail.

43.3.3.1 Factor and Levels

- **Workload:** amount of requests sent for the recommender system.
- **Candidate Items:** candidates are those products that will possibly be recommended. So, this factor limits the number of candidates that the algorithm will be looking for.
- **Neighborhood Size:** controls the neighborhood of users or items. It is commonly used in collaborative filtering.
- **Recommendation list length:** limits the max length of the recommendation list.
- **Alternative Recommendation:** states if an alternative recommendation strategy shall be used, in the case where the recommendation list could not be completed filled.

These input parameters should be adjusted to get the best results. For instance, neighbors' small size can lead to less accurate recommendations, and a big size could cause delays. Finding optimal values is one of the situations that performance evaluation helps to determine.

43.3.3.2 Response Variables

Even though it is imperative to recommend relevant items, the accuracy shall not be the only criterion because it gives only a narrow view about recommendations. For example, long videos on *Youtube* enable more advertisements, which can increase consumption. Therefore, when two videos are similar in relevance, recommending the longest one would bring more benefits.

RSE has eleven metrics (or response variables) split into five groups, as presented in Table 43.1. They measure the behavior of an algorithm concerning some perspective.

Similar metrics tend to be in the same group. If we choose throughput, for example, it is unnecessary to use the response time. They are not the same but equivalents. However, there can be uncorrelated metrics in the same group, like user coverage and novelty. When different groups are involved in the same evaluation, it tends to be more comprehensive.

Table 43.1 Response variables divided into groups

Group	Response variables
Accuracy	RMSE
Decision-support	Precision, recall, F1
User-centered	User coverage, catalog coverage, novelty
Ranking	nDCG
Performance	Response time, response time with queue, throughput

43.3.3.3 Load Distribution

Care must be exercised to ensure that the distribution of users, items, and ratings are not biased. An offline evaluation must simulate the real situation in the best possible way. Also, there is a chance that the algorithm achieved an excellent performance because the test data was suitable for him. Cross-validation is executed on test data to reduce these mishaps, splitting it into train and test sets.

More specifically, users in RSE are divided into five equal-sized groups. One slice is active per replication, and the workload is a subset of it. In personalized recommendations, train and test must be done for the same user. Therefore, user ratings are also divided into groups: history and validation. So in the training step, only the past is available. At the time of the experiment, it would be as if the validation ratings did not exist yet.

43.3.3.4 Evaluation

While evaluating, the appraiser learns new insights that can lead to another test with a different setup. To expand the possibilities, RSE has two evaluation modes:

- **Multifactorial** 2^k : many factors can be used in this approach, but only two variation levels per factor are allowed. It simplifies the experiment, and we can still get reliable conclusions.
- **Multilevel**: in this mode, only one factor is allowed, but it can have many levels. Thus it is possible to get an idea of which are the best values for that factor. It can be used to adjust the parameter or even combined with the multifactorial evaluation.

43.3.4 Reports

After finishing the evaluation, measures will be stored in a specially designed database and further viewed. This module will retrieve data, make some calculations, and present it in a graphical way to help appraisers get conclusions. Currently, RSE has bar charts, line charts, pie charts, and histograms.

43.4 Experimental Results

In order to demonstrate the applicability of RSE, we proposed a use case demonstration of it. For this, we analyze some existing recommendation algorithms using RSE, show how it can be used, and highlight the usefulness of its outputs.

43.4.1 Environment

For experiments, a computer with Ubuntu 16.04 was used. It has an Intel Core2 Quad CPU Q9550 ×64 processor,

2.833 Ghz, and 4 GB of RAM. No concurrent task was running along with the experiment. The warm-up function was executed before each test to stabilize the JVM.

43.4.2 Test Data

Two datasets were used to perform the experiments for validation. RSE can handle with other types of datasets. The two datasets will be briefly outlined.

The authors of [11] show that the movie domain is the most frequent in recommender systems research. So, the *MovieLens10M* [12] dataset was chosen because it is a good example with plenty of data available that has the key characteristics expected from a recommender system. This dataset has 10,000,054 ratings, 95,580 tags, 10,681 movies and 71,567 users. Ratings are on a 1 to 5 scale, and each user rated at least 20 movies. Furthermore, it was incremented using the *MovieLens Tag Genome* [13], which has 1,128 different tags and about 11 million relevance scores applied to movies.

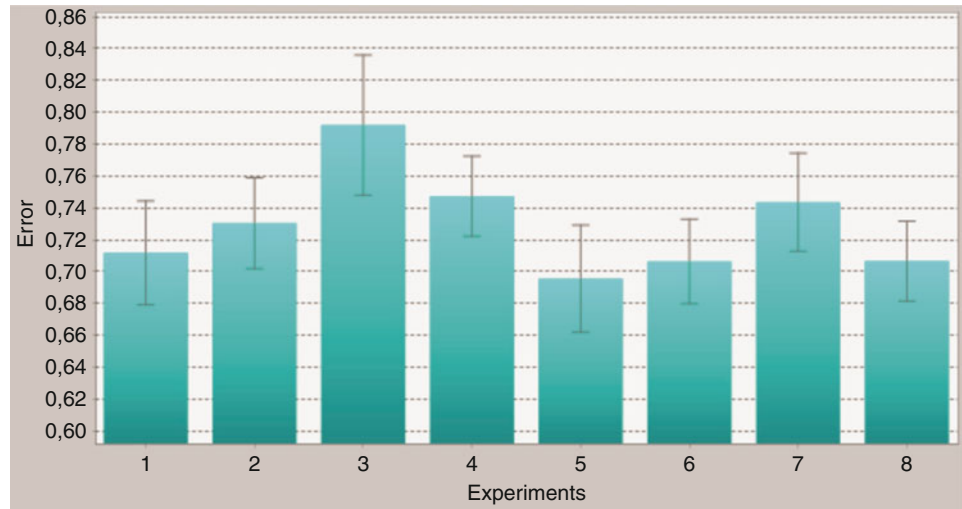
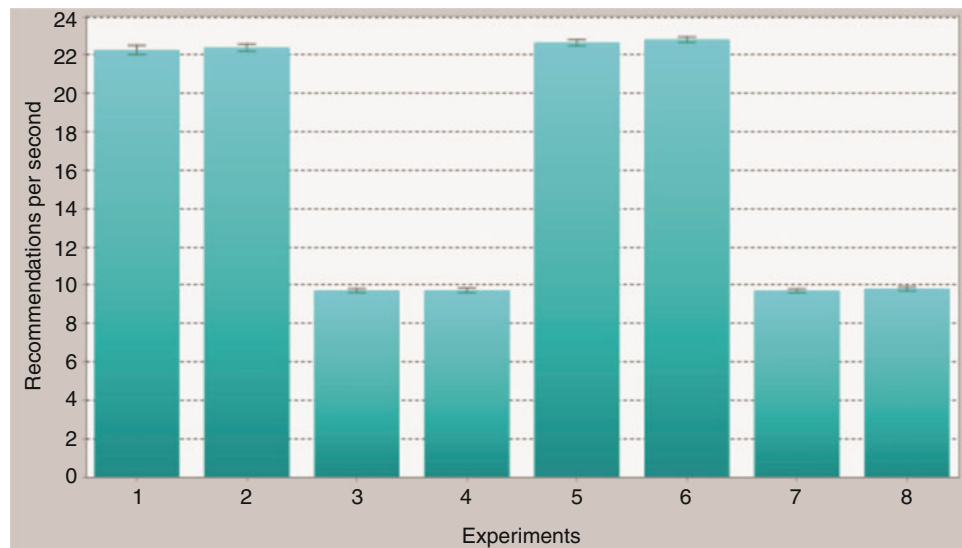
In general, the test data comes in text files. RSE provides methods to facilitate their manipulation and to place it inside the database (e.g., *importData()*, *splitRatings75()*, and many others). The appraiser needs to choose the desired parameters and call the appropriate function.

43.4.3 Evaluations

The test evaluated the performance of *item-item collaborative filtering*. We admit with 90% confidence that the measured outputs had at most 14% of variation around its average value. This computation resulted in 13 repetitions of each experiment. The *relevance* specification allows items to become part of the recommendation list only if they were equal or higher than the assigned value. The complete setup for this test is shown in Table 43.2.

Table 43.2 Setup for the assay

Response variables	Accuracy (via RMSE error)
	Throughput
	Catalog coverage
	User coverage
Factors (1; –1)	A. Neighborhood size (5; 25)
	B. Candidate items (30; 100)
	C. Recommendation list length (5; 25)
Algorithm	Item-item collaborative filtering
Database	MovieLens 10 M
Confidence interval	90%
Ceiling amplitude of CI	14%
Replicas	13
Workload	2500 users per test
Relevance	3.5

Fig. 43.3 RMSE with zoom**Fig. 43.4** Throughput

Figures 43.3 and 43.4 present results for RMSE (root-mean-square error) and throughput in a graphical form. The zoom provided by RSE was used to facilitate the interpretations. Although the confidence interval was overlapping in many tests, experiments numbers 1, 5, 6, and 8 reached the best precision compared to number 3, which was the less accurate one. The best throughputs occurred on tests numbers 1, 2, 5, and 6, reaching around 22 recommendations per second. The experiments 3 and 7 (Table 43.3) achieved nearly 90% of catalog coverage. For user coverage, test 3, 4, 7, and 8 had values among 93% and 94.5%.

Another conclusion from Fig. 43.4 is that it was mainly affected by candidate items (factor B) because variations in throughput are in its change of level. The experiments 1, 2, 5, and 6 were performed with 30 candidates, and the other ones with 100 candidate items. Indeed, this is confirmed by Table 43.4, showing 99.94% of influence. It also shows the influence of factors on other variables.

The choice of the best combination depends on a lot of the application goals. An option that maximizes all of the response variables will not always be available. Table 43.5 summarizes the best combinations found for each metric. If catalog coverage of 74.2% and user coverage of 87.3% were acceptable, experiment 1 could be chosen because it has the best precision and throughput results. If coverage criteria were the most important one, test number 7 would be an option, even though it will cause significant loss of throughput and perhaps some accuracy. Therefore, the combinations given by experiments numbers 1 and 7 were chosen as possible implementations, and they are shown in Table 43.6.

43.5 Conclusion

RSE was built to facilitate the performance evaluation of recommender systems. It has a broad set of metrics, and it is based on statistical concepts to provide reliable conclusions.

Table 43.3 Experimental results for the assay

Exp	Factors			
	A	B	C	
1	1	1	1	
2	1	1	-1	
3	1	-1	1	
4	1	-1	-1	
5	-1	1	1	
6	-1	1	-1	
7	-1	-1	1	
8	-1	-1	-1	
Response variables				
Exp	RMSE	CI%	Troughp.	CI%
1	0.712	9.165	22.271	2.177
2	0.730	7.818	22.390	1.712
3	0.792	11.110	9.714	2.029
4	0.747	6.733	9.730	2.355
5	0.695	9.669	22.654	1.354
6	0.706	7.506	22.814	1.185
7	0.743	8.280	9.705	1.971
8	0.706	7.103	9.805	1.954
Exp	Cat.Cov.	CI%	Usr.Cov.	CI%
1	74.241	0.944	87.385	0.573
2	36.031	1.300	87.462	0.991
3	89.329	0.682	94.154	0.582
4	68.433	1.423	94.385	0.530
5	73.594	1.031	86.462	1.003
6	35.851	1.447	87.077	0.560
7	88.586	0.695	93.539	0.548
8	67.902	1.331	93.615	0.535

Table 43.4 Factor influence for the assay

Factors	Response variables			
	RMSE	Throu.	Cat.Cov.	Usr.Cov.
A	30.12	0.03	0.02	0.96
B	38.02	99.94	37.33	98.80
C	4.90	0.01	57.71	0.13
AB	4.31	0.02	0.00	0.00
AC	0.00	0.00	0.00	0.02
BC	22.23	0.00	4.94	0.02
ABC	0.43	0.00	0.00	0.06

It was developed in modules aiming at ease of extension. Our evaluation demonstrated its simplicity and effectiveness.

The future works are in the sense of expanding the possibilities. Thus more factors, response variables, different correlations, and other divisions between history and test ratings can be added. We want to enable the appraisers to add a multifactorial design with more than two levels. Plugins may be provided to facilitate even more the evaluation. This

Table 43.5 Best experiments for each response variable

Response variable	Experiments
RMSE	1, 5, 6, 8
Throughput	1, 2, 5, 6
Catalog coverage	3, 7
User coverage	3, 4, 7, 8

Table 43.6 Details of tests 1 and 7

Factor	Exp. 1	Exp. 7
Neighborhood size	5	25
Candidate items	30	100
Recommendation list length	5	5

way, algorithms could be imported from the well-known recommendation frameworks.

One could use part of the framework to build a performance evaluation in other contexts rather than recommender systems. To this end, it would be necessary to implement new algorithms, factors, and response variables, but the statistic logic and the output of results would be reused. RSE tries to hide the complexity of creating an evaluation, so we hope it could be useful for developers or appraisers to solve their issues.

Acknowledgements This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) [Finance Code 001].

References

1. M. Nadeem, E.B. Allen, B.J. Williams, A method for recommending computer-security training for software developers: Leveraging the power of static analysis techniques and vulnerability repositories, in *2015 12th International Conference on Information Technology New Generations*, (2015), pp. 534–539
2. J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: a survey. *Decision Support Systems* **74**, 12–32 (2015)
3. S. Trewin, Knowledge-based recommender systems. *Encyclopedia of Library and Information Science* **69**(Supplement 32), 180 (2000)
4. B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in *Proceedings of the 10th International Conference on World Wide Web*, (ACM, 2001), pp. 285–295
5. L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems. *Phys. Rep.* **519**(1), 1–49 (2012)
6. D. Kowald, S. Kopeinik, E. Lex, The TagRec framework as a toolkit for the development of tag-based recommender systems, in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, (ACM, 2017), pp. 23–28
7. A. Said, B. i. Alejandro, RiVal: A toolkit to foster reproducibility in recommender system evaluation, in *Proceedings of the 8th ACM Conference on Recommender Systems*, (ACM, 2014), pp. 371–372
8. M.D. Ekstrand, M. Ludwig, J.A. Konstan, J.T. Riedl, Rethinking the recommender research ecosystem: Reproducibility, openness,

- and Lenskit, in *Proceedings of the fifth ACM conference on Recommender Systems*, (ACM, 2011), pp. 133–140
9. A. Dayan, G. Katz, N. Biasdi, L. Rokach, B. Shapira, A. Aydin, R. Schwaiger, R. Fishel, Recommenders benchmark framework, in *Proceedings of the fifth ACM conference on Recommender Systems*, (ACM, 2011), pp. 353–354
 10. R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling* (Wiley, New York, 1990)
 11. D.H. Park, H.K. Kim, I.Y. Choi, J.K. Kim, A literature review and classification of recommender systems research. *Expert Syst. Appl.* **39**(11), 10059–10072 (2012)
 12. F. Maxwell Harper, J.A. Konstan, The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* **5**(4), 19 (2016)
 13. J. Vig, S. Sen, J. Riedl, The tag genome: Encoding community knowledge to support novel interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)* **2**(3), 13 (2012)

Abstract

The internet has the role of sharing data of the most diverse natures in a fast and accessible way. Some data carries position information within the geographic space. One of the ways used to disseminate this type of data is the Geographic Information System on Web (GIS-Web), which provides a visualization of the data based on its position in the geographic area. Currently, GIS-Webs are present in several scenarios, with an intention to plan or predict the future perspective for rated scenario. Given this, a search was performed in the bibliography on applications of this type of solution, a search was made in 3 repositories with a final result of 15 articles that provided information about which tools or techniques were used, and in which scenario the availability through GIS-Web was used. The illuminated studies of the growing search for the visualization of georeferenced data using elements in 3 dimensions to represent, for example, buildings and sewage or water pipes. The great use of libraries and APIs was identified when making the data available, but the biggest job is in the treatment of raw data that are particular to each application environment. This work presents a systematic literature review, where it is possible to identify possible strategies to be adopted in future works that use georeferenced data.

Keywords

City planning · Map based · Visualization · Geographic data · Geographic information system · Web GIS · Heat map · 3D GIS · 2D GIS · Georeferenced data · Predictability

L. L. G. Duarte (✉) · A. D. de Souza
Federal University of Itajubá, Itajubá, Brazil

44.1 Introduction

Currently, one of the great means of propagating information is the World Wide Web (WWW), this is due to the great ease of making available any type of data, as well as its high reach and ease of access anywhere in the world.

The geographic information system (GIS) is a computational system that allows collect, storage, retrieval, transformation and visualization of geographic data [1]. GIS also has the ability to address spatial relationships between geographic objects and reveal spatial correlations [2, 3].

In view of the need to provide data that are interconnected with some geographic information, such as a city, a neighborhood, a company, platforms were created in a web environment capable of making available and even manipulating georeferenced data. These platforms are called geographic information systems in a web environment (GIS-Web), which according to [4] is a computer system made available through a web platform, used to understand the facts and phenomena that occur in geographic space.

The Internet, with all its versatility, can provide the availability of this type of information in a clear and interactive way, an essential feature in geographical information [1].

For [5] the distribution of data related to geographic space has been increasingly applied in order to understand and elucidate different issues in different scenarios.

It is important to understand the current scenario of applications that aim to provide geographic data, as well as which techniques are most used when storing and disseminating this type of data before start to develop a solution. More than 11% of software projects are canceled before even delivering the first expected result to the customer [6], because of that it is necessary to obtain the greatest amount of information about what is most current in the area in which it is intended to develop a system.

In order to answer the aforementioned questions and to provide a knowledge about actually GIS-Web technologies, a systematic review of the current literature was carried out to identify research using the availability of data through GIS-Web. This study aims to help researchers to understand the current state of the art as well as to find points to be explored in future research.

The paper is organized as follows. Section 44.2 presents the research methodology applied to conduct this work. The Sect. 44.3 analysis the articles found and selected for the research. The Sect. 44.4 discusses the results obtained based on the research questions found in the Sect. 44.2. The Sect. 44.5 presents the conclusion of this work.

44.2 Research Definition

The current work brought together articles published in the last 10 years, from 3 different main repositories for Science Computer, ACM, IEEE and Scopus.

The methodology used in this work is described in the next sub-sessions.

44.2.1 Objective

This systematic review has as main objective to assist future research in understanding which methods are currently used for visualization through the web of georeferenced data.

1. *Population*: The population chosen for this work were articles that use strategies through web platforms to provide georeferenced data.
2. *Intervention*: Search for tools, techniques and scenarios where this type of strategy was adopted.
3. *Comparison*: Not applicable.
4. *Result*: This work seeks the union of the techniques used in the provision of georeferenced data through the web, in order to guide future developments of application that require this strategy.

44.2.2 Research Questions

The research questions for the current study are:

- Q1: What strategies are currently being used in the process of making georeferenced data available in the web?
- Q2: What scenario currently have this type of strategy applied?

On the question Q1, it aims identify which strategies (tools, libraries, API's) are being used to make available data

containing geographic information. The question Q2, aims to identify which scenarios currently GIS-Web is present and to identify gaps and opportunities for futures applications.

44.2.3 Data Collection

The research data was extracted from searches in 3 repositories, they are ACM, IEE and Scopus. The following search term was applied to each of the repositories, with the necessary formatting required for each of the repositories.

- IEEE: (“Full Text & Metadata”: “city planning”) AND (“Full Text & Metadata”: “data visualization”) AND (“Full Text & Metadata”: “GIS” OR “Full Text & Metadata”: “geo-data” OR “Full Text & Metadata”: “geographic data”) AND (“Full Text & Metadata”: “map” OR “Full Text & Metadata”: “map based”) AND (“Full Text & Metadata”: “technology” OR “Full Text & Metadata”: “tool”).
- ACM: (“city planning”) AND (“data visualization”) AND (“GIS” OR “geo-data” OR “geographic data”) AND (“map” OR “map based”) AND (“technology” OR “tool”).
- Scopus: (“city planning” AND “data visualization” AND “GIS” OR “geo-data” OR “geographic data” AND “map” OR “map based” AND “technology” OR “tool”).

The search returned 58 articles (7 from the ACM repository, 27 from the IEEE repository and 24 from the Scopus repository). Among these, the following exclusion (EC) and inclusion (IC) criteria were applied:

- IC-1: Articles that mention the type of strategy used may be selected.
- IC-2: Articles that mention the applied scenarios may be selected.
- CE-1: Articles that do not mention the strategy used will not be selected.
- CE-2: Articles whose scenario is not mentioned will not be selected.
- CE-3: Articles that do not qualify for making data available on the web will not be selected.

After applying the inclusion and exclusion criteria, a result of 15 articles was obtained. Among them, we obtained 2 systematic reviews of the literature, and 2 more articles were obtained within the topic addressed and which were included. The articles used are listed in Table 44.1 and will be discussed in the Sect. 44.3.

Table 44.1 Articles included in the systematic review

ID	Author, reference	Year	Title	H5 index
1	X. Wang et al., [7]	2016	Traffic and transportation smart with cloud computing on big data	35
2	M. Dubey et al., [8]	2019	Predicting biker density at Bikeshare Station Intersections in San Francisco	16
3	X. Huang et al., [9]	2016	TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data	70
4	M. Picozzi et al., [10]	2013	Traffic visualization: Applying information visualization techniques to enhance traffic planning	8
5	S. Davies et al., [11]	2009	Non-intrusive techniques for enhancing decentralized data storage with strategic GIS visualization	Not classified
6	W. Zeng and Y. Ye, [12]	2018	VitalVizor: A visual analytics system for studying urban vitality	28
7	Y. Ran et al., [13]	2010	Research and implementation of three-dimensional visualization based on Internet	7
8	J. Li and N. Chen, [14]	2014	Geospatial sensor web resource management system for smart city: Design and implementation	Not classified
9	X. Li et al., [15]	2015	XEarth: A 3D GIS platform for managing massive city information	11
10	M. Billger et al., [16]	2017	In search of visualization challenges: The development and implementation of visualization tools for supporting dialogue in urban planning processes	22
11	M. Breunig et al., [17]	2020	Geospatial data management research: Progress and future directions	38
12	E. Stefanakis, [18]	2017	Web Mercator and raster tile maps: Two cornerstones of online map service providers	5
13	A.-H. Hor et al., [19]	2018	A semantic graph database for BIM-GIS integrated information model for an intelligent urban mobility web application	29
14	Obie et al., [20]	2018	PedaViz: Visualising hour-level pedestrian activity	Not classified
15	Li et al., [21]	2015	Traffic management and forecasting system based on 3D GIS	Not classified

For the articles selected for this research, information about the tools and techniques used to make geographic data available was extracted, then a snowball technique was applied in order to include articles that can contribute to the elaboration of the work.

From the identifiers (ID) present in the Table 44.1 a conceptual map was created in order to facilitate the understanding in the division of articles between the types of data they present, as well as the area of their application. This map is represented in Figure 44.1, and for the sake of understanding simplification, articles that deal only with data in 2 dimensions are allocated to the 2D direction, and articles that have both techniques are allocated for the 3D direction, the two systematic reviews are to the left of the main item for citing works from both categories.

44.3 Articles Analysis

One of the main points that leads to a broader analysis of the GIS-Web scenario is the use of visualization in 2 dimensions or 3 dimensions (2D or 3D). All 15 articles present some type of 2D data availability, but 8 of them present a 3D model together, which makes visualization and user interaction more attractive.

The next sessions discuss the selected articles summarizing their purpose and techniques applied to prepare and provide the georeferenced data.

44.3.1 2D Data

In the traffic scenario in cities, we can divide the work of [7–10, 20] in 2 categories. Flow analysis and forecasting.

In their works [7, 8] adopt the machine learning mechanism to treat the data. The data are collected from various different sources such as vehicle GPS, traffic events and parking information. The main objective is forecast future events. At work [7] they use machine learning so that in a big data scenario data can be filtered. The filter apply objective is to identify within a large volume of data, which data are used in your work. Dubey et al. [8] used the machine learning techniques for forecasting accidents of rented bikes in San Francisco Bay – US. This information collected was showed through a heat map.

Already [9, 10] in their work only displayed the data that were obtained on urban traffic. The method of making the data available in both cases is done through free APIs provided by Google.

In the [7] the data was collected from GPS placed in taxis in order to create a graph model of the city streets. After that the data was available through a map with information of the average speed and average time to go of a point to another. It also provides a visualization of variation over time, to assist in studies of flow in the roads.

Picozzi et al. [10] work with a proposal to analyze data at intersections in the city. For this, the data are obtained quantitatively by vehicle meters located at the traffic lights

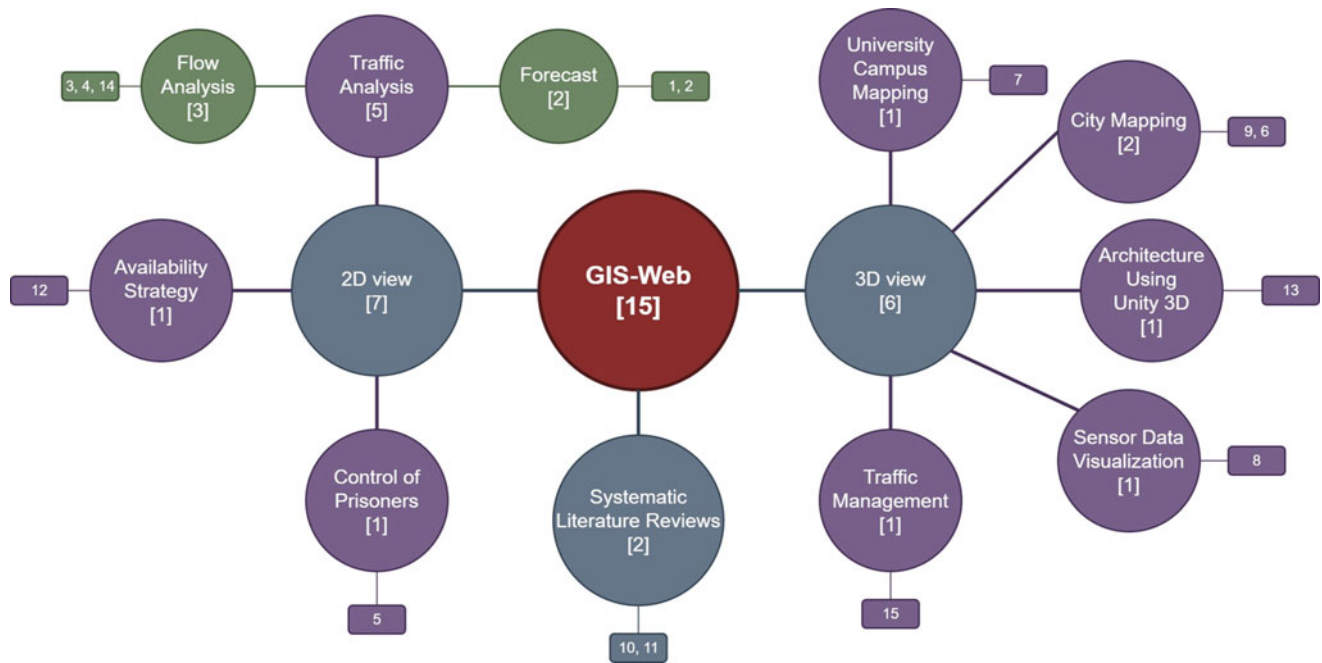


Fig. 44.1 Conceptual map articles and application scenario

of the roads. The generated visualization allows the user to see intersection traffic level (if there are too many vehicles or not), with the possibility of filtering by period, thus being able to analyze the behavior of the roads on different days and times.

In the work of [11] which deals with the development of a GIS-Web to assist in the control of prisoners who are on parole and monitored by electronic anklets. This work proposes the data consumption by a map server (MapServer) that would use georeferenced information to create layers in Shape File format. That data was made available on web pages with the MapServer default map. Within the application, filters are available to find individuals and obtain their location, as well as list all the prisoners data (crimes committed, name, address, parole supervisor name). This system aims to facilitate the management of officers over their supervisors, since an officer may have several prisoners under his responsibility.

Stefanakis [18] used the vector tile strategy. In this strategy the full image of the layer is divided in multi parts, to reduce the size and allow caching of data. This strategy according to [22] is extremely advantageous in the scenario of web applications. Is able to significantly reduces the file size and the request response time. This strategy can optimize the render time of the map in the web pages.

In the work of [20] they present a visualization for sensor data disposed at strategic points in the city of Melbourne in Australia. That's sensors were responsible to collect flow of people in each point. In this work the purpose was to create a 3d interactive map to attend 4 objectives. The first objective

was only counting the people flow in a specific time range. The second objective was to identify the people accumulation pattern in the local of sensors was installed. The third objective was comparing the people accumulation in a specific time range cross with the registered temperature with sensor. This objective intended to identify the people increase in the places owned the lowest temperatures. The fourth objective was to verify the people flow variation in specials dates, holidays and cities events. The solution created used the D3.js library to create and render interactive graphs. Those graphs, was utilized to show the count people, heat maps to show the temperature data collected. The Mapbox was used to render the city map in the web page.

44.3.2 Mixed Data (2D and 3D)

Considering the maps using 3 dimension the works found show this strategy applied in different context. Using the 3D and 2D data [12] uses a different strategy for the elaboration of the visualization. Using APIs to game development [12], created a 3D view of an area of 319 km² in the Netherlands. Using this strategy, the authors was gain optimization in the render of the 3D objects in the map, considering that tools for game development are developed considering the optimization in rendering 3D objects. To create a categorization to the buildings in the map, the authors used colors to classify given it utilization (houses, offices, leisure areas). The Mapview API allows 2- and 3-dimension layers to be overlapping. This allow to view the buildings inside your terrain delimitation

categorized by color by your accessibility level. The streets of the city also display in the map categorized by the accessibility level.

As [12, 13] developed a solution with three-dimensional buildings over the bi-dimensional map. The case study was the campus of the Chinese University of Geoscience and the buildings was mapped and this data was stored in geographic database. The API presented by [13], was developed by the authors of work using Transparent Data Encryption (TDE), provided by ArcGIS to protect sensitive data. In the result of the work the authors mention that the tests were efficient in the 2D and 3D render data.

In a distributed environment [14], was developed a web application to display the sensors data collected. This work was thought and developed by the researchers using the Model View Controller (MVC) pattern to display the georeferenced data. The solution adopts the Open Geospatial Consortium (OGC) prerogatives. The data is sent for the sensors informing their position in real time and them can be dynamics or statics. The dynamic sensors can alter their position over time, the most dynamics sensors are in satellites orbiting the earth. This is a particular environment where the strategies applied to static and little modified data cannot be used. The system behind the application was developed considering all these requirements, but the work not specify the strategy adopted to treat this data type, only show the web application develop strategy.

In the [15] a system was developed to provide a geospatial data visualization from the city of Shenzhen in China. Similar to [12, 15] show data from roads, terrains limitations and buildings, but in this work, them also show the pipe networks data under the city. The utilized tools to manipulate and create 3D models for this solution was C++, OpenGL and ActiveX. To make the data available on the web the authors only used HTML, JavaScript (Js) and C#. In the results of the work they report that this solution is capable to substitute the traditional solutions what using only 2D data models.

Hor et al. [19] was developed a platform to available geographic data, using Unity 3D, a develop game platform. In this work they extracted the data to produce a informative graphs from the classes of 3D objects of GIS and Building Information Model (BIM). The BIM data type, is utilized to map and manage information of building project besides proving the geometric of build and their textures. The data extracted of those 2 data types, was used in a JSON format to create graphs using the Neo4j library. To provide the 3D data, they extracted the necessary information about the geometry and create a Scene Layer Package (SLP). The SLP is a file format optimized for viewing large amounts of data in 3D format. The SLP file is made available through ArcGIS and provide in the web through web services of Unity3D platform.

In the work of [18], a strategy is shown that has been widely used in the display of geospatial data in two dimensions. The strategy presented is the tile map, where different parts of an image are joined based on individual requests from the parties through the web. This strategy according to [22] is extremely advantageous in the scenario of web applications, as it significantly reduces the size of the file, consequently the time to receive data and the decrease in network traffic.

To provide a tool for decision making about traffic city [21], developed a quick and intuitive tool to assist this task. This solution receive data from different sources about the traffic of the city. This data was provided by manage high-ways companies, GPS devices inside taxis, buses and private vehicles capable to send geographic information about their position. To develop this solution [21], used a rendering engine WebVR. That is allow provide the collected data in 3D using virtual reality (VR). This work presents a subdivision of the systems present in the solution:

- Basic traffic information management subsystem.
- Dynamic traffic information management subsystem.
- Dynamic traffic network analysis subsystem.
- Auxiliary subsystem for planning decision making.
- 3D traffic geographic information subsystem.

The 3D traffic geographic information subsystem is responsible for making the data collected by the sensors interactively and processed in their respective subsystems. An important piece of information raised in the discussion of the work is the system's ability to identify abnormal data and remove it from processing since the data collection systems are susceptible to failure. The work also mentions the devices that support the tool and they are PC, PDA, cell phone and tablet.

44.3.3 Systematic Reviews

Billger et al. [16] and Breunig et al. [17] elaborated in their works systematic reviews searching for techniques to provide geospatial data using all strategy types not limited to web. Nevertheless, theirs works can provide some works citing web visualization. Billger et al. [16], present in their work techniques to provide questionnaires data about urban planning. The data was collected by questionnaires about areas of urban planning and after processed, this data was provided in web through maps.

Breunig et al. [17] elaborated a work about the integration of geospatial data in GIS, yet they cited about strategies used to using GIS in the web platform. The strategy mentioned by them consists in partitioning the georeferenced data into small parts, this strategy generates a set of data blocks called Vector Tiles. As the strategy cited by [18],

this strategy also improves the performance of the systems making the bandwidth consumption for sending this data to be optimized. Another method cited is Discrete Global Grid Systems (DGGs), a system created by OGC to provide a model for use of conceptual model and compliance [23] to treat geospatial data.

44.4 Discussion

This session discusses the research questions mentioned in the Sect. 44.2.

A. *Q1: What strategies are currently being used in the process of making georeferenced data available in the web?*

The works found in this research show us a growing tendency to use three-dimensional data to provide geospatial data visualization. From the studies present in this work that used three-dimensional visualization of the data, we can see that this kind of data needs specific processing and tools to make it possible to provide this data in the web. There are several tools and libraries that can assist to making geospatial data available as OpenGL, ArcGIS and WebVR (this now discontinued. WebXR substitute is actually use and meets the specifications of the World Wide Web Consortium – W3C).

In some works, the machine learn was applied to provide a forecasting about the future of the scenario. This show us the data georeferenced data goes far beyond what the data itself is capable of providing us. It combined with other knowledge areas, such as machine learning, deep learning and statistics, provide a tool capable to make the tasks of decision making easier.

B. *Q2: What scenario currently have this type of strategy applied?*

The most different scenarios can receive this type of strategy. The found scenarios in this work were, traffic management [7–10, 20], traffic management and forecasting system [21], control of prisoners on parole [11], display of building data for a given scenario [12, 13] and display real-time sensor data [14].

44.5 Conclusion

With the results obtained by this work, it is possible to notice the tree-dimensional data is the most used strategy to provide a interactive visualization about the geospatial data. This strategy allows more complete strategic vision. This allows the decision maker to make the most of what the data can provide to take strategic decisions. View data about that has already occurred can encourage the decision maker to forecast the possible problems that may encountered in the

future. However, if the tool can make this future analysis. the decision maker will have more possible problems that he could not forecast or would take a long time to list all the forecasting generated by machine.

As presented, there are several ways to treat and make georeferenced data available through a web platform. In a web scenario, the elaboration of the architecture, the storage of the large volume of data received and the processing of raw data is the harder step to provide a good visualization data. These tasks become more difficult when the Big Data concept is present and the solution intends to provide a real time visualization of this data. After the data processing the solutions can use existing APIs to render the data in the desired format using interactive maps or graphs. The APIs presented in this systematic review are not the only ones that can be used, currently several APIs are developed and improved every day.

We can relate the different scenarios presents in this work with the smart cities concept. This concept has been increasingly researched in the computer science. Severus techniques and tool were developed and will be developed to provide a good and interactive visualization of geospatial data, with the intention of reducing processing costs and optimizing the task of those who need this type of data on a daily basis.

References

1. V.C.O. Souza, Y.B. Castro, M.M.V. Paula, M.M.L. Volpato, Demarcação: A proposal to support the mapping of coffee areas using citizen science, in *Proceedings of the XIV Brazilian Symposium on Information Systems, ser. SBSI'18*, (Association for Computing Machinery, New York, 2018). [Online]. Available: <https://doi.org/10.1145/3229345.3229367>
2. G. Câmara, Representação computacional de dados geográficos, in *Banco de dados geográficos*, ed. by M. A. Casanova et al., (MundoGEO, Curitiba, 2005), pp. 11–52
3. P.R. Fitz, *Geoprocessamento sem complicação* (Oficina de textos, 2018)
4. M.d.F. de Pina, S.M. Santos, Basic concepts of geographic information systems and cartography applied to health (2000)
5. M. Polidoro, M.V.F. Barros, Methodological proposal for the development of geographic information system in web environment (GIS-Web) applied to tourism. *Ar@cne Revista Electrónica de Recursos en Internet sobre Geografía y Ciencias Sociales* **133** (2010)
6. B.G. Tavares, M. Keil, C.E.S. da Silva, A.D. de Souza, A risk management tool for agile software development. *J. Comput. Inf. Syst.*, 1–10 (2020)
7. X. Wang, Z. Li, Traffic and transportation smart with cloud computing on big data. *Int. J. Comput. Sci. Appl.* **13**(1), 1–16 (2016), publisher: Technomathematics Research Foundation
8. M. Dubey, A.P. Ortiz, R. Agrawal, A.G. Forbes, Predicting biker density at Bikeshare Station Intersections in San Francisco, in *2019 IEEE Global Humanitarian Technology Conference (GHTC)*, (2019), pp. 1–7
9. X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, C. Zhang, TrajGraph: A graph-based visual analytics approach to studying urban network

- centralities using taxi trajectory data. *IEEE Trans. Vis. Comput. Graph.* **22**(1), 160–169 (2016)
10. M. Picozzi, N. Verdezoto, M. Pouke, J. Vajus-Anttila, A. Quigley, Traffic visualization: Applying information visualization techniques to enhance traffic planning, in *GRAPP 2013 IVAPP 2013 – Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications*, (2013), pp. 554–557. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84878146214partnerID=40md5=3f7e484ade0ef6ee781ee1b3ba1f377a>
 11. S. Davies, R. Lamb, N. Odhiambo, Non-intrusive techniques for enhancing decentralized data storage with strategic GIS visualization, in *2009 WRI World Congress on Computer Science and Information Engineering*, vol. 2, (2009), pp. 379–383
 12. W. Zeng, Y. Ye, VitalVizor: A visual analytics system for studying urban vitality. *IEEE Comput. Graph. Appl.* **38**(5), 38–53 (2018)
 13. Y. Ran, K. Zheng, X. Liu, Research and implementation of three-dimensional visualization based on Internet, in *2010 18th International Conference on Geoinformatics*, (2010), pp. 1–4
 14. J. Li, N. Chen, Geospatial sensor web resource management system for smart city: Design and implementation, in *2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, (2014), pp. 819–827
 15. X. Li, Z. Lv, J. Hu, B. Zhang, L. Shi, S. Feng, XEarth: A 3D GIS platform for managing massive city information, in *2015 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, (2015), pp. 1–6
 16. M. Billger, L. Thuvander, B. Wästberg, In search of visualization challenges: The development and implementation of visualization tools for supporting dialogue in urban planning processes. *Environ. Plann. B. Urban Anal. City Sci.* **44**(6), 1012–1035 (2017), publisher: SAGE Publications Ltd
 17. M. Breunig, P. Bradley, M. Jahn, P. Kuper, N. Mazroob, N. Rösch, M. Al-Doori, E. Stefanakis, M. Jadidi, Geospatial data management research: Progress and future directions. *ISPRS Int. J. Geo Inf.* **9**(2) (2020), publisher: MDPI AG
 18. E. Stefanakis, Web Mercator and raster tile maps: two cornerstones of online map service providers. *Geomatica* **71**(2), 10 (2017)
 19. A.-H. Hor, G. Sohn, P. Claudio, M. Jadidi, A. Afnan, A semantic graph database for BIM-GIS integrated information model for an intelligent urban mobility web application. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **IV-4**, 89–96 (2018). [Online]. Available: <https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-4/89/2018/>
 20. H.O. Obie, C. Chua, I. Avazpour, M. Abdelrazek, J. Grundy, T. Bednarz, PedaViz: Visualising hour-level pedestrian activity, in *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction – VINCI '18*, (ACM Press, Växjö, Sweden, 2018), pp. 9–16. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=3231622.3231626>
 21. X. Li, Z. Lv, J. Hu, B. Zhang, L. Yin, C. Zhong, W. Wang, S. Feng, Traffic management and forecasting system based on 3D GIS, in *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, (IEEE, Shenzhen, China, May 2015), pp. 991–998. [Online]. Available: <http://ieeexplore.ieee.org/document/7152585/>
 22. A.C. Robinson, U. Demšar, A.B. Moore, A. Buckley, B. Jiang, K. Field, M.-J. Kraak, S.P. Camboim, C.R. Sluter, Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *Int. J. Cartography* **3**(sup1), 32–60 (2017). [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/23729333.2016.1278151>
 23. Open Geospatial Consortium, Ogc 15-104r5: As topic 21: Dggs, <http://docs.opengeospatial.org/as/15-104r5/15-104r5.html>, August 2017. Accessed on 06/23/2020

Marina Lapenok, Anna Lozinskaya, and Vasilisa Likhacheva

Abstract

The article describes the process of intelligent modeling of pedagogical situations using artificial neural networks, built on the basis of the analysis of cognitive patterns of human information processing, allows the development of effective decision support systems and forecasting learning systems. The experience of creating neural network systems is presented: to predict the success/failure of a student's project activities with the development of recommendations for selecting a perspective project task; to predict student decision to attend/skip classes based on student personal qualities, aims and lesson schedules.

Keywords

Artificial neural networks · Cognitive profiles · Intelligent systems · Pedagogical problems · Programming in Python

fuzzy conditions. Modeling is based on a system analysis of essential (in the context of research) elements and the connections between them, and the synthesis of a network of information transformation units in the form of directed graphs. Usually, the system is built in layers, with a special role played by the ways of organizing layers and connections of network units (neurons).

Currently, in the scientific literature and the media there are many reports of successful experience in using neural networks to create expert predictive systems. The need to use neuroscience in pedagogy has been expressed by many scientists (Moskvin V. A., Eremeeva V. D., Khrizman T. P.). At the end of the twentieth century, it became possible to realize more widely the achievements of neuropsychology and neurobiology in pedagogical practice.

Artificial neural networks to a certain extent simplify the processing of information by natural biological networks of neurons. The architecture of both natural and artificial networks is built on neurons. Many types of neural network architectures have been developed (Feed Forward Networks, Recurrent Networks, Hopfield Network, Kohonen Networks, Hamming Network, etc.) [1, 2]. The following general aspects of neural networks can be distinguished: the cells (neurons) of the network have constant or dynamic roles (input/hidden/output); connections with static or dynamic scales are established between the cells, working in forward/backward passages; neurons are controlled by signal values, which can be regulated by connection weights or activation signal values.

There are two modes of functioning of the neural network – training (settings) and working. The training of the network is necessary to select the weights of the connections that make it possible to achieve localization of excitation at the output layer of neurons with an acceptable error. To train networks, various algorithms are used (backpropagation of an error, tracing, adaptive resonance, fuzzy logic, genetic, fractal, etc.) [2]. In the process of training the network with

45.1 Introduction

The development of an intelligent network model is a solution to cognitively complex problems of identifying and analyzing structural elements, as well as a set of properties and relationships of elements that are significant in the context of target landmarks.

Intelligent modeling is currently mainly implemented on the basis of the development and use of artificial intelligence networks – multilayer neural networks that allow simulating the linear and nonlinear behavior of complex systems in

M. Lapenok (✉) · A. Lozinskaya · V. Likhacheva
Yekaterinburg, Russia
e-mail: lapyonok@uspu.me; alozinskaya@uspu.me;
v.b.lihacheva@uspu.su

the error backpropagation algorithm (with a teacher), many iterations are implemented with a certain step, during which the weights of the connections are selected and built in a certain order to provide the necessary structure of neuron interconnections. During training, ideal (reference) values of the pairs <inputs -outputs> are used.

The cognitive processes of perception and processing of information by a person, reflected in the achievements of the Theory of Attention and Working Memory, the Theory of Mental Representations, the Differentiation-Integration Theory of Development of Thinking, are the basis of simulation models of artificial neural networks and combinations of learning algorithms.

Human systems thinking is based on an associative mechanism: the perceived information is compared with the mental models of knowledge available in the brain and the cognitive pattern that is closest in content is found, new information refines the cognitive model of the network or rebuilds this model [3]. Cognitive networks of mental models of knowledge are constantly evolving, the landscapes of cognitive spaces are changing. It is associative memory – memory by content – that is modeled by many neural networks, since it can be used in multimedia systems and ensures stable operation with incomplete/damaged input/output parameters [1–3].

45.2 Results and Discussion

Typical pedagogical tasks (in the context of developing an intelligent network model) are teaching, classification, forecasting, decision making. In pedagogy, neural networks are developed for the implementation of adaptive learning systems (reinforcing [4]/individual [5]); categorization (learners/pedagogical problems); development of individual recommendations (nutrition/physical activity/schedule of classes/rhythm of the training load, taking into account preferences and psycho-physiological characteristics of health); predicting the success of activities (educational/cognitive/project [6]; diagnostics of educational achievements of students.

Modeling of a neural network is implemented at the following main stages: (1) formalization of the problem (formation of descriptive characteristics of the initial data and typical samples of solutions); (2) model construction – the selection of discrete units of the system and significant relations (formation of a matrix of input parameters and output samples of solutions); (3) the choice of the type of neural network for solving a specific problem and the method of training it (determining the structure of the network and training algorithms); (4) development of an intelligent network and its configuration (programming signal processing

by the network); (5) use of the network (using the network to solve relevant problems, self-learning of the network).

The process of intelligent modeling of solving pedagogical problems using neural networks is inextricably linked with cognitive issues, which can be attributed to the following groups:

1. detailing the pedagogical experience for manipulating the elements of a problem task/situation (analysis of the cognitive experience of an expert teacher; cognitive analysis of typical characteristics/signs and relationships);
2. development of a matrix model of a pedagogical task (analysis of cognitive models of thinking and behavior of students; cognitive analysis of typical characteristics/attributes and relationships, samples of solutions/situations);
3. development of a neural network and mathematical learning models (cognitive analysis of a simulation mathematical model and cognitive models of human activity);
4. application of the network (cognitive analysis of new pedagogical tasks/situations).

Detailed pedagogical empirical experience is based on the description and analysis of typical examples of a pedagogical task or situation: groups of characteristics and relationships, interrelationships and patterns that most significantly affect the specific result of solving the problem/development of the situation are distinguished. This activity requires the expert educator to have the cognitive skills to differentiate the elements of the whole; correlate, compare, generalize and organize elements; to highlight the core characteristics for the subsequent integration of elements into the model of the network system.

The creation of a matrix model of a pedagogical task/situation is based on the specification of the selected characteristics in the course of modeling the cognitive profiles of students and samples of typical solutions. Lists of characteristics (features/parameters) of the problem pedagogical task are created (these will be the input parameters of the network), which can be heterogeneous, their values can be assessed on a dichotomous scale (0, 1) or quantitatively (natural numbers). The most important input data of the intelligent network for supporting the solution of pedagogical problems are the cognitive characteristics of students (parameters of cognitive profiles): a cognitive network of initial subject knowledge and experience of activity (not/indicative/basic/deep); style of thinking (analytical, pragmatic, creative); cognitive styles (coding information, processing information, posing and solving problems, building a picture of the world); characteristics of memory and attention; comfortable pace of work; resistance to interference; reliance on teachers [7]. The input parameters of the network can include other charac-

teristics of the subjects and conditions of the pedagogical process (purposefulness, leadership qualities, the ability to work in a team, sociability, tolerance, temperament, active interaction with the environment, motivation, knowledge of equipment, knowledge of research methods, time and number of attempts to complete the task, the quality of task completion).

The development of a neural network and mathematical learning models requires a cognitive analysis of the pedagogical task and the correlation of its results with the structure and configuration of various types of neural networks – for an adequate choice of network architecture and functioning algorithms. Application of the network to solving relevant pedagogical problems contributes to its learning on new examples and development.

The experience in designing data mining systems includes the following examples of creating neural network systems capable of predicting:

1. the decision of students to visit/skip training sessions (depending on the personal characteristics and target attitudes of students, identified through a survey), in order to optimize the planning of the educational process and the organization of social and psychological support for students [8];
2. the success/failure of the project activities of students (depending on their personal characteristics and the portfolio of educational achievements) in order to create a methodology for the selection of parameters of project activities that ensure success [9].

The artificial neural networks can be visualized in the form of a directed graph, the vertices of which will correspond to neurons, and the arcs connecting the vertices will correspond to synaptic connections or weights (Fig. 45.1).

Example 1 Prediction of attendance by students of training based on their personal qualities, goals and lesson schedules.

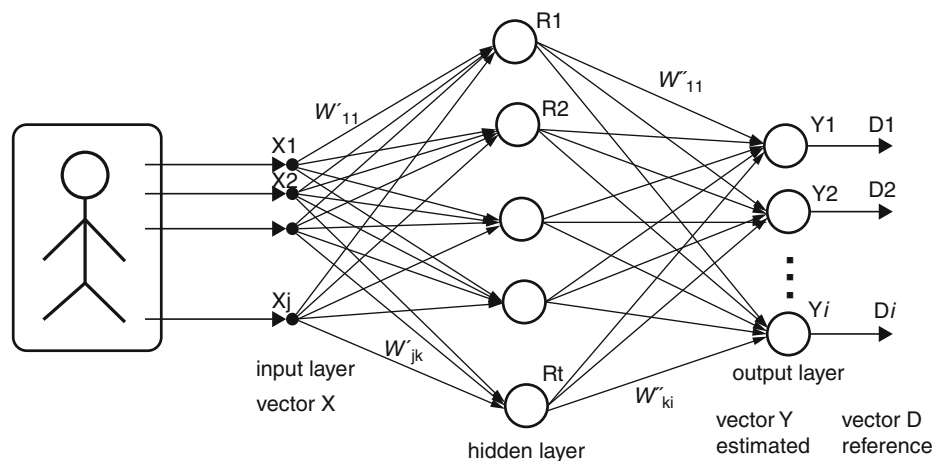
To form the input vector X and the output calculation vector Y (and, accordingly, the output reference vector D), their dimensions, and the contents of each component, it is necessary to analyze the studied area, relying on typical samples of its implementation. The decision of students to attend training sessions in a situation prevailing at a particular time depends on their personal characteristics, which can be identified using questionnaires. Among personal characteristics, it is advisable to consider, for example, the following:

- self-organization (ability to resist weaknesses);
- psychotype (extrovert, introvert, ambivert);
- the type of temperament (choleric, sanguine, phlegmatic, melancholic);
- responsibility (or irresponsibility) of the student in relation to the implementation of public assignments;
- the prevalence of the desire to acquire new knowledge in the learning process over the desire to acquire a document on the availability of this knowledge as a result of study (or vice versa);
- a student’s assessment of the importance of their own time in terms of the appropriateness of time’s transport costs, depending on the expected duration of the classes (for example, only one pair in the class schedule or the presence of “windows”);
- concern over the state of one’s own health in terms of the intention to attend (or not attend) training sessions;
- a sphere of subject interests (programming, computer games, theoretical foundations of computer science);
- the aim is to combine study with professional activity (or not), etc.

These are the input parameters of the vector X used for calculation by a neural network.

In the output vector D , the results of attendance by the students surveyed by training sessions (for example, in specialized disciplines) taken from group journals of the dean’s

Fig. 45.1 The scheme of the neural network



office should be encoded. The empirical data obtained as a result of the questionnaire and from the data of the dean's office are reduced to a dichotomous scale of yes to no and are used to design and train a neural network.

Example 2 Prediction of the results of project activities of students.

We will observe how the teacher makes the choice “perspective” for students of project assignments. The term “perspective” project assignment means that the topic of the assignment will interest the student, that the student has a range of knowledge and skills sufficient to complete the project assignment, that the student's personal characteristics (determination, perseverance, ability to analyze, etc.) will allow him to cope with the intended volume of analytical and/or experimental work.

The teacher forms a record in which he fixes the name, the name of the group; then, as a rule, it finds out the interests and activity of the student (for example, does the student engage in circles in the framework of further education or electives, whether he participated in competitions and how successfully); listens to the wishes of the student in relation to the scientific field within which he intends to carry out the project; gets acquainted with the results of educational activities for previous periods of training; finds out how diligent (for example, regarding homework) and discipline (for example, did he miss classes for no good reason, did he receive comments for his bad behavior). You can also take into account sociability (person-person, person-nature, person-sign, person-technician), the type of temperament of the student and other data that determine the characteristics of a person and, therefore, affect the likelihood of systematic purposeful work (collective or individual) on the project task.

As a result, the teacher accumulates several parameters characterizing the personality of the learner and the history of his educational activity, having processed which with the help of his knowledge and pedagogical experience, the teacher draws a conclusion about the likelihood of successful completion of the task, and then formalizes and presents the student with the content of the project activity. These are the initial parameters - input for human analysis or input for calculation by a neural network, defining vector X.

In the output vector D, the possible results of the project activity should be encoded.

To create neural network forecasting systems that solve pedagogical problems, special software was developed using a high-level cross-platform Python programming language. During the development of a computer program, it became necessary to import standard modules and additional libraries, such as a Pickle module for serializing and deserializing objects, a Random module for generating pseudo

random numbers, for the exact calculation of the exponent of a number, etc.

45.3 Conclusions

An analysis of our own experience in designing predictive neural networks and implementing them in the Python programming language made it possible to come to certain conclusions:

- The effective development of an intellectual model requires a sufficiently large base of initial data, in connection with which it is necessary to conduct targeted pre-testing/questioning of a certain category of students (for example, first-year students) on the developed groups of psychological and pedagogical issues and track the effectiveness of their activities over time (in the context of the study of pedagogical tasks). In other words, the allocated cognitive profiles of students should be compared with the activity histories, which will make it possible to form pairs of input and output (exemplary) network parameters and develop effective teaching examples.
- The system of input and output parameters of the intellectual system should be developed based on the signs of associative thinking: cause and effect, proximity in time or space, similarity or commonality, contrast or opposition, belonging to a class, belonging to the whole, modality.
- The solution of pedagogical problems of clustering and forecasting is fully provided by a 1–3-layer neural network trained by the error back propagation algorithm.
- Along with the use of simulators and neural network frameworks (TensorFlow, Brain, Primat, Nest, Playground, Neuroph), an intelligent neural network system can be implemented in the Python programming language using the Tkinter, Pickle, NumPy libraries.

In conclusion, we note that the development of neural networks to support the solution of pedagogical tasks based on the analysis of the parameters of cognitive processes and profiles, allows us to use the unique advantages of network computing methods and tools to create an adaptive and SMART educational environment.

References

1. M.T. Jones, *Artificial Intelligence Programming in Applications/-Translated from English Osipov A.I.*, 2nd edn. (DMK Press, Moscow, 2013), 312 p
2. A.B. Barsky, *Neural Networks: Recognition, Control, Decision Making* (Finance and Statistics, Moscow, 2004), 176 p
3. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach/-Translated from English*, 2nd edn. (Williams, Moscow, 2006), 1408 p

4. M. Ausin, H. Azizoltani, T. Barnes, M. Chi, Leveraging deep reinforcement learning for pedagogical policy induction in an intelligent tutoring system, in *Proceedings of the 12th International Conference on Educational Data Mining*, (2019), pp. 168–177
5. M.F. Caro, D.P. Josyula, J.A. Jiménez, Multi-level pedagogical model for the personalization of pedagogical strategies in intelligent tutoring systems. *Dyna (Medellin, Colombia)* **82**, 185–193 (2015)
6. M.V. Lapenok, A.M. Lozinskaya, Cognitive approach to teaching intellectual modeling of pedagogical problems, in *Proceedings of the 7th International Scientific-Practical Conference*, (Yekaterinburg, 2019), pp. 248–254
7. M.A. Kholodnaya, *Cognitive Styles. On the Nature of the Individual Mind*, 2nd edn. (SPb, Peter, 2004), 384 p
8. M.V. Lapenok, I.V. Rozhina, N.G. Tagiltseva, V.D. Likhacheva, Neural network prognostic systems for research and solution of pedagogical problems, in *Proceedings of ISERD 174 International Conference*, (New York, 16–17 October 2019), pp. 13–17
9. M.V. Lapenok, O.M. Patrusheva, S.A. Hudyakova, Using neural network mathematical models to solve pedagogical problems, in *Proceedings of the International Scientific Conference “Digitalization of Education: History, Trends and Prospects”*. *DETP 2020*, vol. 437, (Atlantis Press), pp. 22–26

Immersive Virtual Reality and Its Use in Developing Empathy in Undergraduate Students

Éder Estrada Villalba and Fausto Abraham Jacques-García

Abstract

In recent years, Immersive Virtual Reality (IVR) has taken on great relevance in the academic and scientific world, especially in relation to the development of empathy. Virtual Reality (VR) allows taking perspective of others through virtual environments and, through this experience it is plausible to think we can promote empathic capacity. This work is part of an ongoing doctoral research and presents the initial findings about the topic and problem statement through a diagnosis and literature review on VR for the development of empathic responses. Findings include reassertion about the general potential of VR, continuous usage of the Interpersonal Reactivity Index (IRI) to quantify empathy and, the lack of rigorous evidence to prove VR could be the ultimate empathy machine.

Keywords

Developing empathy · Educational technology · Emerging technology · Empathic capacity · Empathy machine · Immersive virtual reality · Literature review · Perspective taking · Virtual reality · Virtual reality perspective taking

last 5 years is unable to give enough and rigorous evidence to confirm if this technology can be used or apply it to improve learning outcomes or if it is or could be a more effective or meaningful method. This article presents the initial part, a diagnosis and the literature review, of an applied research work that seeks to answer questions such as: Is Taking Perspective in Virtual Reality (PTVR) a more significant learning experience to develop empathic capacity; to imagine how the other feels from his point of view through virtual embodiment? Can the level of development of empathic response through PTVR be greater than that achieved through traditional Perspective Taking (PT) methods?

The diagnosis and literature review have the purpose of validating a theoretical and practical problem that arises from a need to base the institutional Educational Technology strategy of the “Zonas VR” of the Tecnológico de Monterrey that are being deployed from its emerging technology for education laboratory since 2019. This work is part of a broader study inserted with in the doctoral program in Innovation in Educational Technology of the Autonomous University of Querétaro whose objective is to develop an emerging pedagogy model to design learning experiences with Immersive Virtual Reality (IVR) to develop empathy.

46.1 Introduction

The effects of the use of modern Virtual Reality (VR) in education are still not clear and the empiric studies from the

É. Estrada Villalba (✉)
Humanities and Education School, Tecnológico de Monterrey,
Santiago de Querétaro, Qro., México
e-mail: eder.villalba@tec.mx

F. A. Jacques-García
Informatics School, Universidad Autónoma de Querétaro,
Santiago de Querétaro, Qro., México

46.2 Virtual Reality and Empathy

Virtual Reality (VR) is a communication medium that makes virtual experiences feel real and it has been used by the military and medicine for training and simulations purposes since 1960, but VR has also become fertile ground to evaluate social and psychological dynamics in academic settings [1]. Journalists use VR to create immersive news, placing their audience at the center of the stories and from other’s perspectives [2], educators use VR for situated learning, putting their students in an environment that otherwise would not be possible in a presential instruction setting and, psychiatrists

use VR to mitigate the negative effects of psychological trauma [3]. For a VR environment to be perceived as real and for the mind and body to take it seriously, it needs to “feel” as much real as possible. Previous and recent studies on this topic have shown that there are three phenomena that directly impact the sensation of reality in a virtual environment: immersion, presence and embodiment, which refer to design aspects and technological capacities that enable a virtual experience [3–14].

Empathy is fundamental for positive human interaction and the construction of a social fabric based on solidarity and compassion, as well as the development of a culture of peace; The role of empathy is decisive in all sciences, it is considered an integral part of education and necessary in pedagogical processes to help with the social and emotional development of students [15, 16]. Promoting empathic responses and the concept of empathy itself is problematic, and for that reason it is suggested that this question should be placed at the center of the debate about whether VR can promote or develop empathy [17] and, despite the fact that some case studies on immersive journalism suggest that this technology can promote it [17] there is no firm evidence to declare that VR can be more effective than other less or non-immersive or technological mediated methods to develop empathic responses.

46.3 General Objectives

Support the approach to the research problem and present the literature review on the application of IVR in formal learning processes to promote the development of empathy through interview and survey of teachers and students from Tec de Monterrey, as well as previous academic and scientific studies published between 2015 and 2020. The purpose of the complete doctoral research work is to develop an emerging pedagogy model to design meaningful learning experiences with IVR to promote the development of empathy in a formal higher education context.

46.4 Method

To carry out the initial diagnosis, qualitative methods were used, specifically a scale survey applied to university students and a semi-structured interview to professors from Humanities faculty to obtain approaches from experience and knowledge, explanations and reasoning on the subject of study. For the construction of the literature review, we used the method described by O. L. Londoño, L. F. Maldonado, & L. C. Calderón [18] that refers to two major phases: (1) Heuristics: search and compilation of information sources and (2) Hermeneutics: analysis and interpretation of information.

46.4.1 Methodology

Data collection instruments: the scale survey seeks that the participants count and value some of the approaches from their experience and knowledge and, that they explain their reasoning freely. While the semi-structured interview poses a set of open questions formulated in a specific order, of which follow-up and probing questions can be included to explore a topic in greater depth [19]. To carry out this study, letters of informed consent (digital and printed) were presented to the participants, clearly stating the purposes of the research, their participation as voluntary, interviews were going to be recorded (in audio) and their right to have the results of the study at the end of it.

Data treatment: regarding the survey, the instrument was applied until reaching saturation, that is, enough information to allow observing patterns, constructing conjectures, making correlations and inferences. The collected data were processed by content analysis (explicit and latent) through coding, categorization and integration of these. This is a technique for the systematic analysis of oral or written content that can be used in materials such as diaries, letters, courses, dialogues, reports, books, articles and other linguistic expressions [20].

Iterative heuristic process: from preparation to selection of documents for the literature review. The sources of information were delimited by relevance, importance and reputation in the field of study and the academic and scientific community. Databases consulted were ProQuest Dissertations & Theses Global, ScienceDirect, SciVal, Scopus, WOS, EBSCO Education, JSTOR, ProQuest Central, ProQuest ERIC, Redalyc, RITEC, SAGE Premier, SciELO, Taylor and Francis Journals. Advanced queries were made with various key words, in different combinations, both in Spanish and English: Virtual Reality, Virtual Reality and Education, Virtual Reality and Empathy, Immersive Virtual Reality, Virtual Reality and Perspective Taking, to mention the most important ones. After the initial search, only 87 publications were found to be of relevant for the study according to their titles, abstracts and keywords. After the previous selection, introductions of all publications were reviewed to verify which ones fully met the criteria resulting in the final sample of 34 articles.

Hermeneutical process: from interpretation to publication of the literature review. This is the critical analysis of the selected information, contextualizing and evaluating previous results and achievements, establish knowledge gaps and aspects to be addressed in future studies, identify trends and assess limitations within the research topic.

46.5 Results

The diagnosis showed that from the students' perception (96 students participated): the usability of technology is seen as a motivational element rather than a formative one (65%), they recognize that the technological capacities of presence, immersion and embodiment can help in developing empathy and, they intuit and speculate that it can increase empathy through "putting oneself on the other's shoes" (31%), finally, they ignore the impact VR can have on education, and in particular on development empathic capacity (44%). From the teachers' perception (seven teachers participated): two teachers directly mentioned the development of empathy and one, taking of perspective as a key factor to take advantage of VR, referring to the immersive capacities that would allow to put oneself in the place of the other, all professors recognize that the technological capacities of presence, immersion and embodiment can be important elements to trigger a learning process but, they also ignore the impact VR can have on education, and on developing empathic responses. The sample for the literature review was comprised from academic and scientific articles (85%) as well from degree theses or dissertations (15%) from 16 different countries; in the Table 46.1 we present a synthesis of the most notable findings:

In the Table 46.2 we breakdown the interest, development and advances over the time period between 2015 and

2020 about the research topic, through this examination we observed an interesting perspective and a prospective of possible future studies:

It is worth noting that throughout this years, educational technology researchers have shown various levels of interest in this topic from an exploratory and general perspective on the use of VR in education.

46.6 Discussion

Despite the inherent conceptual difficulty of empathy, previous studies have a constant framework, those maintained the same instrument to measure empathy, the Interpersonal Reactivity Index (IRI) created by Mark H. Davis in 1983. Although we know that PT it is more effective than only having information about the other's situation to feel empathy [21], that VR experiences have the potential to increase PT towards other specific individuals [24] as well as empathic concern [23] and, that this suggests that TPRV could be more effective than other methods [21].

However, it is imperative to solve one of the biggest limitations we have encountered: the inadequate quantification of empathy; the IRI has not been revised and fully used as only cognitive empathy has been discussed while we also have emotional and somatic empathy [27]. Research on the application of VR in educational contexts is still incipient but

Table 46.1 Most notable findings from previous studies

Previous results and conclusions	
Empathy is not only a trait and it can increase or decrease [15, 21].	
PT is more effective than just providing data to make people feel empathetic [21].	
VR simulations offer a promising training environment of professionals [22].	
VR has potential to increase engagement and consequentially, increase empathy towards specific individuals [23]; increase could be regulated by the sense of immersion in the virtual environment [24].	
Immersion is directly related to presence, and to how "real" the mediated environment feels; tracking, stereoscopy and field of vision are key factors [9].	
Embodiment and body transfer in VR can affect the perception of oneself and their treatment of others [25].	
It is suggested that PTVR could be more effective in developing empathy than traditional ones [21].	
Current gaps and limitations	
Experimental designs from previous studies have been limited in number and diversity of participants.	
Quantification or measure of empathy has been problematic and not adequately addressed.	
Increase in empathy in some of the studies may have been due to the exposure or novelty of the technology.	
There is insufficient evidence to confirm the suggestion about PTVR being a more effective method to develop empathic responses than non-immersive methods.	
We could not find records of PTVR to develop empathy in formal higher education contexts.	
Most of the studies focuses on increasing motivation, engagement and academic performance.	
The technical variation of VR equipment and its constant technological improvement limits the number of participants and the replicability.	

Table 46.2 Interest about the research idea

2015	Chris Milk's provocation "VR the ultimate empathy machine?" [26]; first studies and empirical works.
2016	Empirical explorations were carried out measuring VR as media in contrast to other technological tools.
2017	Studies focused on the variables of immersion in VR that could possibly result in increase in empathy.
2018	Critical elements emerged: quantification of empathy, need of longitudinal studies and, examination the effect of exposure and novelty.
2019	Other fields of study became involved and ethical concerns were included in the discussion ("forcing" empathy through VR).
2020	Methodological limitations persist and hinder the generation of concrete conclusions.

the studies to which we have access have revealed potential and possible benefits that they can contribute to learning process or outcomes, hence, in the last years this topic have become more prevalent within the discussion of new pedagogical approaches (emerging pedagogies) and ideas to take advantage of those technologies.

46.7 Conclusions

While a significant percentage of students think that VR can be a relevant educational resource, they state it in terms of making classes or learning contents more dynamic, interesting, motivating or didactic; right now, they see the technology behind VR more as a motivational element rather than a formative one. We did not ask explicitly if PTVR could be a more meaningful form of learning, however, they mentioned ideas such as “putting oneself in the shoes of the other”, “being a protagonist”, “experiencing other realities” which means they see (speculative) potential VR through presence, embodiment and immersion can help increase empathy through “putting on the other’s shoes”. Between students and professors, there is no clear or uniform understanding of the possibilities that VR opens up in terms of ethical and civic education, we found repeatedly responses like “I don’t know”, “I don’t know how”, “I don’t see how” about the use or application of VR in education.

IVR in education is still a very recent topic that deals with an emerging technology that is constantly evolving, which makes clear the need to conduct more (empirical) experimentation and rigorous (scientific) research in this field. This work helps to support and justify the need of more studies and educational experiments related to this topic.

Acknowledgments We appreciate the support from the Faculty of Information Technologies of Autonomous University of Querétaro (UAQ); the National Council of Science and Technology (CONACYT); and the National Office of Faculty Development of Tecnológico de Monterrey.

References

1. D. Markowitz, J.N. Bailenson, *Virtual Reality and Communication*. (Oxford Bibliographies, 2019). <https://www.oxfordbibliographies.com/view/document/obo-9780199756841/obo-9780199756841-0222.xml>
2. A. Jelfs, D. Whitlock, The notion of presence in virtual learning environments: What makes the environment “real”. *Br. J. Educ. Technol.* **31**(2), 145–152 (2000). <https://doi.org/10.1111/1467-8535.00145>
3. L. Sánchez-Laws, Can immersive journalism enhance empathy? *Digit. Journal.* **8**(2), 213–228 (2017). <https://doi.org/10.1080/21670811.2017.1389286>
4. A. Shaskevich, Virtual reality can help make people more compassionate compared to other media, new Stanford study finds. *Stanford News* (2018). <https://news.stanford.edu/2018/10/17/virtual-reality-can-help-make-people-empathetic/>
5. A. van Loon, J.N. Bailenson, J. Zaki, J. Bostick, R. Willer, Virtual reality perspective-taking increases cognitive empathy for specific others. *PLoS One* **13**(8), e0202442 (2018). <https://doi.org/10.1371/journal.pone.0202442>
6. S. Hasler, B. Spanlang, M. Slater, Virtual race transformation reverses racial in-group bias. *PLoS One* **12**(4), e0174965 (2017). <https://doi.org/10.1371/journal.pone.0174965>
7. A. Monje, *Metodología de la investigación cuantitativa y cualitativa: Guía didáctica* (Universidad Surcolombiana, Neiva, 2011)
8. C. Milk, How virtual reality can create the ultimate empathy machine, in *TED Ideas worth spreading*, 2015. https://www.ted.com/talks/chris_milk_how_virtual_reality_can_create_the_ultimate_empathy_machine
9. S. Oh, J.N. Bailenson, G. Welch, A systematic review of social presence: Definition, antecedents, and implications. *Front. Robot. AI* **5** (2018). <https://doi.org/10.3389/frobt.2018.00114>
10. P. Banakou, D. Hanumanthu, M. Slater, Virtual embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Front. Hum. Neurosci.* **10**, 601 (2016). <https://doi.org/10.3389/fnhum.2016.00601>
11. Gallego, *Competencias ética y ciudadana* (documento inédito). Instituto Tecnológico y de Estudios Superiores de Monterrey, Nuevo León, México, 2014
12. E. Behm-Morawitz, H. Pennell, A.G. Speno, The effects of virtual racial embodiment in a gaming app on reducing prejudice. *Commun. Monogr.* **83**(3), 396–418 (2016)
13. J. Herrera, N. Bailenson, E. Weisz, E. Ogle, J. Zaki, Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PLoS One* **13**(10), e0204494 (2018). <https://doi.org/10.1371/journal.pone.0204494>
14. J.J. Cummings, J.N. Bailenson, How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychol.* **19**(2), 272–309 (2016). <https://doi.org/10.1080/15213269.2015.1015740>
15. J.O. Bailey, J.N. Bailenson, D. Casasanto, When does virtual embodiment change our minds? *Presence Teleop. Virt.* **25**(2), 222–233 (2016)
16. K. Kim, L. Bölling, S. Haesler, J. N. Bailenson, G. Bruder, G. Welch, Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar, in *Proceedings of the IEEE International Symposium for Mixed and Augmented Reality 2018* (To appear)
17. K. Stavroulia, A. Lanitis, Enhancing reflection and empathy skills via using a virtual reality based learning framework. *Int. J. Emerg. Technol. Learn.* **14**(07), 18–36 (2019). <https://doi.org/10.3991/ijet.v14i07.9946>
18. M.J. Mayan, Una introducción a los métodos cualitativos: Módulo de entrenamiento para estudiantes y profesionales. Alberta: International Institute for Qualitative Methodology (Qual Institute Press, 2001). <https://sites.ualberta.ca/iqim/pdfs/introduccion.pdf>
19. N.J. Formosa, B.W. Morrison, G. Hill, D. Stone, Testing the efficacy of a virtual reality-based simulation in enhancing users’ knowledge, attitudes, and empathy relating to psychosis. *Aust. J. Psychol.* **70**(1), 57–65 (2018) <https://doi.org/millennium.itesm.mx/10.1111/ajpy.12167>
20. N. Schutte, E. Stilinović, Facilitating empathy through virtual reality. *Motiv. Emot.* **41**(6), 708–712 (2017) <https://doi.org/millennium.itesm.mx/10.1007/s11031-017-9641-7>
21. O.L. Londoño, L.F. Maldonado, L.C. Calderón, *Guía para construir estados del arte* (International Corporation of Networks of Knowledge, ICONK, Bogotá, 2016). ISBN: 978-958-57262-2-2
22. R. Cable, What is virtual reality? And what are its most pressing implications?. *Stanford Humanities Center. Research news* (2019), <http://shc.stanford.edu/news/stories/what-virtual-reality>

23. R. Hassan, Digitality, virtual reality and the 'empathy machine'. *Digit. Journal.* **8**, 195–212 (2019). <https://doi.org/10.1080/21670811.2018.1517604>
24. S.J. Ahn, J. Bostick, E. Ogle, K.L. Nowak, K.T. McGillicuddy, J.N. Bailenson, Experiencing nature: Embodying animals in immersive virtual environments increases inclusion of nature in self and involvement with nature. *J. Comput. Mediat. Commun.* **21**(6), 399–419 (2016)
25. S.J. Ahn, J.N. Bailenson, D. Park, Short-and long-term effects of embodied experiences in immersive virtual environments on environmental locus of control and behavior. *Comput. Human Behav.* **39**, 235–245 (2014)
26. U.N. Ramírez, J.F. Barragán, University students' selfperception on the use of digital technologies for learning. *Apertura* **10**(2), 94–109 (2018). <https://doi.org/10.32870/ap.v10n2.1401>
27. W.B. Owais, E. Yaacoub, Quantifying empathy in virtual reality, An Outline. *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, 2020. <https://doi.org/10.1109/iciot48696.2020.908>

Fangyang Shen, Janine Roccosalvo, Jun Zhang, Yun Tian, Yang Yi, Yanqing Ji, Ashwin Satyanarayana, Xiangdong Li, Ahmet Mete Kok, Annie Han, and Hon Jie Teo

Abstract

This research analyzes how remote learning models are utilized in STEM Education. The E-NEST project developed online teaching models to instruct teaching interns during the unprecedented times of the coronavirus pandemic including mentorships, internships and culturally responsive teaching summer workshops. Based on key findings from data collection and program evaluations from the National Science Foundation Robert Noyce Teacher Scholarship program, a comprehensive online learning classroom was created to teach cultural diversity in STEM Education with modified project management and recruitment approaches. As a result, E-NEST online internships and professional development workshops

were effective and promoted student achievement during the transition to remote learning. The project team learned the functions of online apps to instruct students and gained experience in facilitating online learning classes.

Keywords

Noyce · Online instruction · Digital technology · Culturally responsive teaching · Online project management · Professional development · Internships · Mentorships · Summer workshops · Data collection

F. Shen (✉) · J. Roccosalvo · A. Satyanarayana · X. Li · H. J. Teo
Department of CST & CTTE, NYC College of Technology (CUNY),
Brooklyn, NY, USA

e-mail: fshen@citytech.cuny.edu; jroccosalvo@citytech.cuny.edu;
asatyanarayana@citytech.cuny.edu; xli@citytech.cuny.edu;
hteo@citytech.cuny.edu

J. Zhang
Department of Math & CS, University of Maryland, Eastern Shore,
Princess Anne, MD, USA
e-mail: jzhang@umes.edu

Y. Tian
Department of Computer Science, Eastern Washington University,
Cheney, WA, USA
e-mail: ytian@ewu.edu

Y. Yi
Department of ECE, Virginia Tech, Blacksburg, VA, USA
e-mail: yangyi8@vt.edu

Y. Ji
Department of ECE, Gonzaga University, Spokane, WA, USA
e-mail: ji@gonzaga.edu

A. M. Kok · A. Han
Department of CIS & Math, Borough of Manhattan Community
College (CUNY), New York, NY, USA
e-mail: amkok@bmcc.cuny.edu; yhan@bmcc.cuny.edu

47.1 Introduction

The beginning phase of this E-NEST National Science Foundation (NSF) grant faced pronounced challenges due to the unprecedented event of the coronavirus pandemic affecting the entire country. New York City College of Technology (City Tech) and Borough of Manhattan Community College (BMCC) are located in the epicenter of the pandemic in the United States where universities could not continue to conduct in-person classes due to the sudden onset of a national emergency.

The main goal of this new NSF Noyce grant is to recruit and educate prospective STEM teachers for high need schools in Brooklyn and New York City (NYC) public school districts. Since in-person classes were no longer encouraged, STEM faculty from these universities explored methods and found a new approach to train future K-12 educators. Conceptual framework factors influencing E-NEST remote learning were identified and consisted of project management, internships, mentorships, Noyce summer workshops, online teaching and recruitment. E-NEST faculty developed remote learning professional development workshops for college students. The E-NEST project, in particular, implemented

specific online teaching models emphasizing cultural diversity in STEM Education.

To support the continuation of teaching and learning, an assortment of new online learning tools was utilized. The two main apps, Zoom and Microsoft Teams greatly helped to sustain the quality of the E-NEST project. Other online tools used for professional development workshops and project management meetings were Skype, Cisco Webex, Google hangouts, Google classroom and Blackboard Ultra. These apps were also used for intern and scholar recruitment and summer internships and mentorships.

All of these online apps offer advanced functions to give interns the opportunity to adapt to remote learning effortlessly. This aided in teaching the best practices in cultural diversity in STEM Education. These online tools were used during the Noyce Summer Workshops which have been offered to teaching interns and scholars annually since the first NSF Noyce grant. Although these summer workshops were not able to be conducted in-person, the core components of these workshops were still able to be taught excellently.

The E-NEST project team was able to learn the functions of the Zoom and Microsoft Teams apps efficiently and gain experience in facilitating online learning and project team meetings using all other apps. There were numerous advantages of remote learning which will be further discussed in Sect. 47.3 of this paper. We will continue to develop innovative online learning models which will be applied to support all interns, scholars and teachers who will partake in this NSF Noyce program.

During the E-NEST project, Noyce interns and scholars participate in annual summer programs consisting of STEM Education and cultural diversity professional development workshops. Also, they are involved in mentorships with STEM and Education faculty from City Tech and BMCC. At the end of this NSF Noyce grant period, highly effective STEM teachers will be trained to teach in NYC schools where there has been a teaching shortage, specifically in STEM Education.

The rest of this paper is organized as follows: Section 47.2 reviews the literature for this topic; Sect. 47.3 introduces E-NEST online learning models to train interns and scholars to become exceptional prospective STEM teachers and to conduct meetings remotely with E-NEST faculty; Sect. 47.4 presents project data collection and external program evaluations comprised of surveys and interviews from Noyce students; Sect. 47.5 summarizes the findings of this study and discusses potential guidelines for upcoming research.

47.2 Literature Review

According to [1], the United States has a critical shortage of qualified STEM educators in classrooms, especially in high-need NYC public schools. With uncertified mathematics and science teachers in junior high schools and high schools, students are significantly disadvantaged in learning. Not only recruiting, but retaining STEM teachers in these schools has affected student achievement in several subject areas including computer science and mathematics. The NSF Noyce program at City Tech and BMCC addresses this issue by recruiting K-12 teachers for high-need schools in urban communities.

In [2], these studies have confirmed that teacher turnover rates have had the most impact in high-need communities where students are underprivileged. These teachers often leave these communities to teach in school districts where there is a lower percentage of minority students and higher socioeconomic status. Without competent teachers, student success is difficult to attain in high-need communities. The Noyce program not only recruits, but retains STEM educators by offering supportive and informative training for teachers.

The NSF has developed the Robert Noyce Teacher Scholarship program [3] to assist in recruiting and retaining certified STEM teachers to pursue careers in the United States. The Robert Noyce Teacher Scholarship program seeks to encourage talented Science, Technology, Engineering and Mathematics majors and professionals to become K-12 STEM teachers in high-need local education agencies.

Students are eligible to receive up to a maximum of 2 years of Noyce scholarship support while they study to become certified to teach a STEM subject. The NSF grant being implemented at City Tech and BMCC has greatly helped high-need schools acquire the certified teachers needed for students in all communities to receive a top-quality education in varied subject areas.

In [4–6], this research presents the three-tiered project model, Noyce Explorer, Scholar and Teacher, which was used in the previous successful Noyce project at City Tech and BMCC. Additional improvements have been identified and will be implemented in the current E-NEST project at both City Tech and BMCC campuses where they will support STEM students. This project has gained success in recruiting students enrolled in STEM majors to obtain their degrees and teaching certifications, specifically in Mathematics and Technology fields. The previous Noyce NEST project productively trained 20 new qualified K-12 STEM teachers for high-need communities in the NYC area.

In [7], the pedagogy, culturally responsive teaching is recognized and demonstrates the significance of including this issue in STEM Education. Culturally responsive teaching is essential to create an inclusive classroom where students increase their knowledge and modify their teaching methods. As a result, Noyce workshops including culturally responsive teaching guide students to create an improved classroom culture relating to awareness of race and ethnicity.

Reference [8] discusses how remote learning, specifically the use of videos, has been beneficial during the coronavirus pandemic. The current generation of students are accustomed to the standard of using video in their personal lives and video should be applied more to higher education practices for student success. This research discusses different approaches to use video and technology in the online classroom.

In [9], this research examines online learning across the nation during the beginning months of the pandemic. State education agency policy guidelines by all 50 U.S. states were analyzed. Several areas of consensus were found including cancellation of testing, recommendations to continue some form of remote learning, attention to digital and non-digital options and a fair and appropriate education for all students, including students with disabilities. By providing digital options in schools, diverse learning styles were accommodated and contributed to further student success.

In [10], a collaborative effort to support teacher educators is outlined. This includes supporting the instructional needs and concerns of educators due to the challenges posed by the global pandemic at the university level. The School of Teacher Education members formed a remote learning community and developed an action plan and implemented resources. Suggestions in this research can be used to enhance future considerations related to teacher professional learning and preparation.

47.3 E-NEST Remote Learning

During mid-March 2020, the coronavirus pandemic affected the United States, especially New York City which was the epicenter of this virus. Due to these unforeseen circumstances, City Tech and BMCC universities could not continue to hold in-person classes and transitioned hurriedly to remote learning models.

Several challenges of E-NEST remote learning were identified involving technological, cultural and personal dynamics. As a result, the E-NEST member committee identified key conceptual framework factors that would affect remote learning during this period as shown in Fig. 47.1.

With the first year of the NSF Noyce project just beginning, the E-NEST project team used the above model to modify their project management meetings, recruitment strategies, internships, mentorships and Noyce summer workshops involving culturally responsive teaching. Without being able

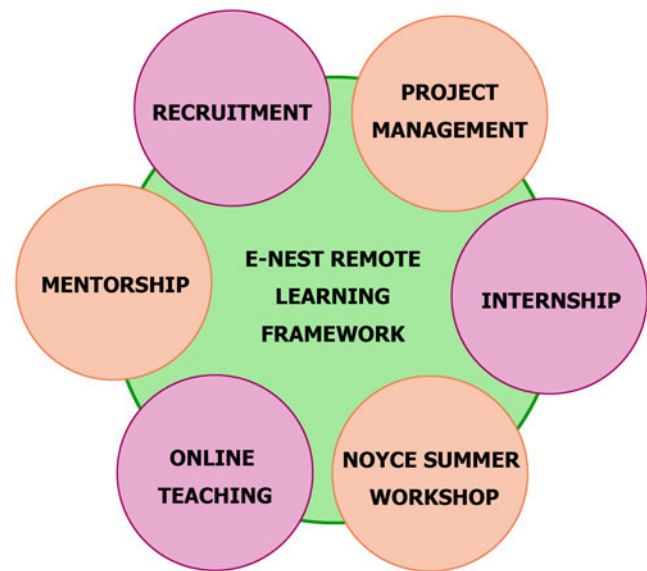


Fig. 47.1 Conceptual framework factors influencing E-NEST remote learning

to hold team meetings, recruitment, mentorships or workshops in-person, the E-NEST project management team re-configured their curriculum to hold all of these project activities remotely for the summer. The two main apps, Zoom and Microsoft Teams were utilized. Other online tools used were Skype, Cisco Webex, Google hangouts, Google meetings and Blackboard Ultra.

47.3.1 Project Management

In-person project management team meetings were not able to be held. Instead, these meetings were conducted using various online tools, emails and phone conferences. Strong leadership from the E-NEST faculty led remote meetings to engage the team and encourage communication. Project management responsibilities included assigning tasks and monitoring and tracking to make sure team members stayed on schedule. Dropbox and Google Drive were also conveniently used to upload and share files.

The E-NEST project management team adapted approaches to organize the Noyce summer workshops; recruit interns for the summer workshops and summer internships; recruit scholars; facilitate internships and mentorships remotely; and utilize online teaching apps and functions.

47.3.2 Recruitment and Mentorships

Noyce intern, summer workshop and scholarship applications were distributed widely through emails. Varied STEM departments each reached out to their cohort of students.

Online seminars and information sessions were conducted online to further advertise the Noyce program to university students.

Materials that students included in their application were: application form, resume, college transcripts, academic recommendation letters and essays. The E-NEST project team met remotely to discuss potential applicants for all Noyce internships, summer workshops and scholarships and conducted student interviews online. Students were selected to participate in Noyce mentorships with professors, internships, summer workshops and scholarships during the summer semester remotely. Online mentorships consisted of individual and group meetings with STEM and Education faculty. Online submission on all student work including project assignments and reports was required. Student field placement records were provided to be completed to log hours and activities completed for internship activities. Remote office hours were provided to potential and existing interns, scholars and summer program students by using waiting room app functions.

47.3.3 Noyce Summer Workshops

The Noyce summer workshops consisted of eight-day four-hour remote learning workshops involving STEM Education subject areas. Topics included culturally responsive teaching, mathematics, computational thinking, introduction to teaching, quantitative reasoning and problem solving, design embedded systems, computer-aided techniques in design and manufacturing and classroom case studies. Hands-on online learning activities were provided where students were engaged in kinesthetic learning.

47.3.4 Benefits of E-NEST Remote Learning

There were several advantages of using the new remote learning models for project management meetings, recruitment and mentorships and Noyce summer workshops.

The E-NEST project team and Noyce students were able to learn the functions of the Zoom and Microsoft Teams apps efficiently and gain experience in facilitating and participating in online learning and project team meetings.

The advanced functions of these apps were valuable for faculty and for students who are already skilled in modern technology apps. Hosts were able to assign other faculty as co-hosts to allow them to use more advanced functions online. Breakout room and chat functions helped better facilitate group and individual discussions. Students were able to ask questions on chat and have faculty address them at a later time during the workshop. Faculty used the chat function to send survey links, website links and reminders.

Online recording of meetings with certain permission was another resource that made information readily accessible for students. Visual aids such as built in tools for screen sharing allowed students and faculty to share information, online labs and PowerPoint slides on hand. Students were able to give presentations online using audio and video features.

Meetings were able to be scheduled ahead of time and recurring meetings were easily emailed to all faculty and students with one link. Attendance lists were automatically generated at the end of the meeting to see which students were present for each workshop. Keyboard shortcuts were also functions used resourcefully.

These apps were convenient for students and faculty to log on from any device and from any location. Making these workshops more accessible for all allows these workshops to host a greater amount of people. Without space as an issue, a larger number of students can be recruited to participate in summer workshops and other meetings.

The Noyce external evaluation was conducted remotely during the Noyce summer workshops by using the breakout room function. The workshop professor helped facilitate students into a separate room to be interviewed by the program evaluator. Survey links were also emailed to summer workshop students to evaluate the Noyce program and give feedback to the evaluator.

47.4 Program Data Collection and Results

During the Summer semester, City Tech and BMCC interns partook in Noyce teaching internships and summer professional development workshops remotely. Students were surveyed online once they completed their internship or summer workshop sessions by interviewing with the program evaluator and/or survey responses.

The findings of this data are based on empirical evidence, which identifies the critical factors of remote learning in the E-NEST program which have contributed to the academic success of Noyce explorers. Based on the results, the respondents generally agreed that E-NEST online internships and summer workshops were effective and promoted student achievement during the transition to remote learning.

Figures 47.2, 47.3, and 47.4 below illustrate significant findings on remote teaching and learning in the Noyce program. In Fig. 47.2 below, the majority of Noyce students strongly agreed and agreed that the online instruction was engaging and enjoyable.

In Fig. 47.3 below, the majority of students strongly agreed that online instruction added higher quality to teaching and learning.

In Fig. 47.4 below, the majority of students strongly agreed and agreed that online instruction empowered them to differ their pedagogical skills.

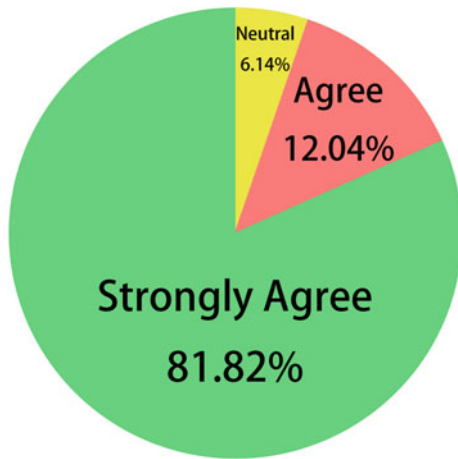


Fig. 47.2 Noyce online instruction engagement Question: The online instruction of the Noyce program was engaging and enjoyable for learners

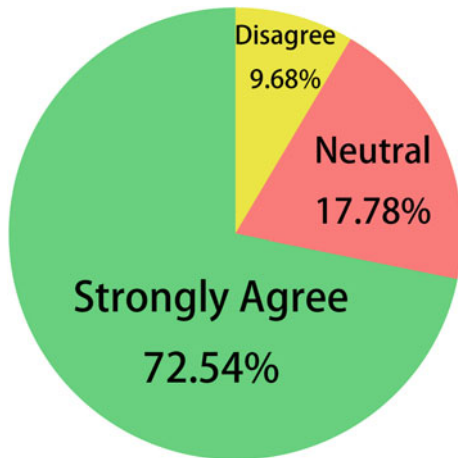


Fig. 47.3 Online teaching and learning Question: I think the use of online learning improves the quality of teaching and learning

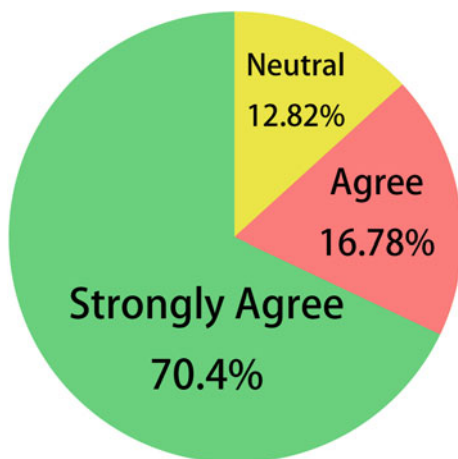


Fig. 47.4 Remote pedagogical skills Question: I believe online learning enables me to vary my pedagogical skills.

This survey data demonstrates that supports for Noyce students including online internships, summer workshop programs and mentoring by STEM and Education faculty have aided students to transition to remote learning successfully. With the E-NEST project officially starting under pandemic conditions, the project team was still able to launch a productive online summer 2020 workshop series along with several teaching internships involving high-need high schools and CUNY professors. Noyce funding has allowed CUNY to continue to encourage talented STEM university students to become K-12 computer science and mathematics teachers.

47.5 Conclusions and Future Work

In this paper, remote learning models that were utilized in STEM Education were analyzed. The E-NEST project modified teaching and project management strategies to accommodate students during the global pandemic of the coronavirus. Modified teaching and project management strategies consisted of issues on recruitment, culturally responsive teaching summer workshops, internships and mentorships.

In addition, program data collection and external program evaluations were presented. Based on key findings from data collection and program evaluations from the NSF Robert Noyce Teacher Scholarship program, a comprehensive online learning classroom was created to teach cultural diversity in STEM Education along with effective recruitment, project management and mentorship strategies. The outcomes demonstrate that the new remote learning strategies during the E-NEST project were constructive and successful.

This research could be applied to many other similar projects nationwide. For future work, we plan to continue to modify remote learning strategies to apply to the new cycle of the E-NEST project. Also, we will collect additional program data and external program evaluations to analyze results and contribute to future STEM Education projects.

Acknowledgments This work is supported by the National Science Foundation (Grant Number #1: NSF 1950142, \$1,444,398, May 1, 2020 – April 30, 2025; Grant Number #2: NSF 1340007, \$1,418,976, January 1, 2014 – December 31, 2019; PI: Fangyang Shen; Co-PI: Mete Kok, Annie Han, Andrew Douglas; Project Manager: Janine Rocco-alvo; Program Assistant: Yanwen Zhu, Kendra Guo, Danping Zhong). The E-NEST project team would like to thank Prof. Gordon Snyder for his help on the project's external evaluation. We also would like to thank all faculty and staffs at both City Tech, BMCC and Research Foundation CUNY who have helped and supported both of our Noyce projects in the past 10 years.

References

1. National Science Board, *Science and Engineering Indicators 2018* (National Science Foundation (NSB-2018-1), Alexandria, 2018)
2. S. Loeb, M. Ronfeldt, J. Wyckoff, How teacher turnover harms student achievement. *Am. Educ. Res. J.* **50**(1), 4–36 (2013)
3. National Science Foundation, *Robert Noyce Teacher Scholarship Program Solicitation NSF 17-541* (2020)
4. F. Shen, J. Roccosalvo, et al., Creating culturally responsive Noyce explorers, scholars and teachers, in *17th International Conference on Information Technology: New Generations*, (Springer Publications, Cham, 2020)
5. F. Shen, J. Roccosalvo, et al., STEM education enrichment in NYC, in *16th International Conference on Information Technology: New Generations*, (Springer Publications, Cham, 2019)
6. F. Shen, J. Roccosalvo, et al., NSF Noyce recruitment and mentorship, in *15th International Conference on Information Technology: New Generations*, (Springer Publications, Cham, 2018)
7. E.S. O’Leary, C. Shapiro, S. Toma, et al., Creating inclusive classrooms by engaging STEM Faculty in Culturally Responsive Teaching Workshops. *Int. J. STEM Educ.* **7**, 32 (2020)
8. D. Lipomi, Video for active and remote learning. *Trends Chem.* **2**(6), 483–485 (2020)
9. J. Reich, C. Buttimer, et al., *Remote Learning Guidance from State Education Agencies during the COVID-19 Pandemic: A First Look* (Massachusetts Institute of Technology, 2020)
10. F. Safi, T. Wenzel, L.A. Trimble Spalding, Remote learning community: Supporting teacher educators during unprecedented times. *J. Technol. Teach. Educ.* **28**(2), 211–222 (2020)

Alejandra Acuña, César Collazos, and Cristian Barría

Abstract

Technology is changing the world, society, and people. For many, there is no conceivable living without it. Its evolution has had exponential growth, that is why many of the futuristic scenarios are no longer so and we are facing an inevitable human transformation; such as, the impact of these new technologies on human life. It is necessary to reflect what the design considerations have been or if there was an analysis that incorporated human and ethical values in its evaluation was carried out. This takes us a little further back to the training of the engineer, who was taught to evaluate technically and economically a technological project, but a methodology to evaluate the scope and impacts that this technological project has were not included in his training. Thus, it is vital to teach how to reflect carefully on the consequences and effects it causes to nature and society. The aim of this research is to offer a purposeful assessment of the incorporation of ethics and human values linked to the training of an engineer in the technological area.

Keywords

Ethics · Software · Design · Computer · Informatics · Technology · Education · Methodology · Human values

A. Acuña (✉) · C. Barría
Centro de Investigación en Ciberseguridad, Universidad Mayor,
Santiago, Chile

A. Acuña · C. Collazos
Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad
del Cauca, Popayán, Colombia
e-mail: alejandra.acuna@umayor.cl; acunaximena@unicauca.edu.co;
ccollazos@unicauca.edu.co

48.1 Introduction

Technology is a factor in the development of humanity and it is possible to recognize the importance of its application and the impact it generates on the quality of people's life and society. With the dizzying penetration of science and technology in all human scenarios, coupled with the inexorable growth of different exponential technologies, artificial intelligence, the Internet of Things, Big data, Robotics, Nanotechnology, virtual/augmented reality, among others [1], techno philosophy was born during the twentieth century, a discipline dedicated to the study of the nature of technology and its social effects. One of its main purposes is to seek the characteristics and understanding of social, economic and political systems to improve the quality of life of people and society [2].

In Chile, the concern regarding the consideration of the social impact of new technologies has gained strength, largely due to the encouragement from different entities to technological innovation and entrepreneurship. It is evident in the new Ministry of Science, Technology, Knowledge and Innovation created in July 2018 by the Government of Chile, to promote science, technology, knowledge and innovation as key transforming agents for the country to achieve development, sustainable and comprehensive, which contributes to improving the quality of life of people and contributes to the development of the territories [3].

The concern of Higher Education Institutions (HEIs) is to analyze how future engineers are being trained in issues of raising awareness and making human and ethical evaluations in the development of technology. Students must learn to reflect carefully on the consequences and effects that technology causes on nature and society.

The problem arises because ethics and human values are not being considered in a systematic and methodological way from the conception of the project and its evaluation in the

social and individual impacts that the product will have once it is finalized and put into use to the different interested parties directly or indirectly. This leads us to review the issue in relation to the education of future professionals in the technology area, specifically engineers in the area of computing and informatics, and how they conceive, create, design and develop technological projects, beyond technical and economic feasibility, the methodologies, and standards used for an appropriate design and software development.

The methodology considers two reviews of the current situation regarding the incorporation of ethics in research of careers. A local review of different Chilean universities that offer careers in engineering, the area of computing and informatics and that in their study programs consider ethics as a subject. The second review considers a survey at the Latin American level of academics who teach subjects associated with software design, to find out the inclusion of ethics as part of the definition in the design of software (see Fig. 48.1).

The purpose of this research is to offer a propositional assessment of the incorporation of ethics and human values linked to the education of an engineer in the technological area, and to promote changes that include a systematic methodology that covers from the conception and design of a technological project until its implementation and use in society, to assess the impact and repercussions on people's lives.

The article presents a vision of works related to ethics in engineering education; the codes of professional ethics of the technology area of different international organizations and

Value Sensitive Design, a design methodology to take into account and incorporate human values throughout the design process, is a systematic attempt to include values of ethical importance in design [4]. The results of the review in Chile on the subject of ethics in the study programs and the results of the survey at the Latin American level on the inclusion of ethics in the subject itself. It continues with the analysis of the results and ends with the conclusions and consideration of future work.

48.2 Works Related to Ethics in High Education

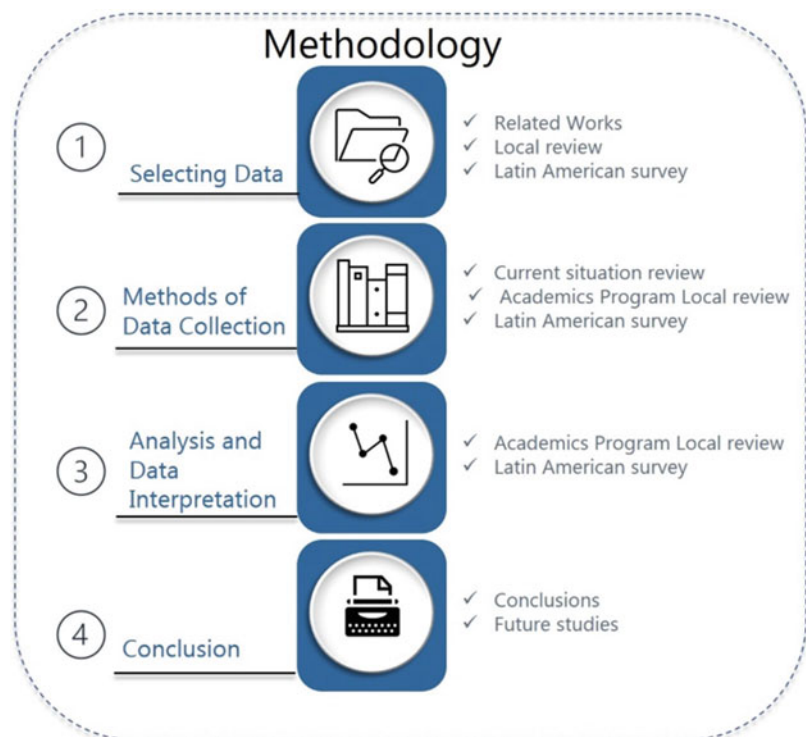
48.2.1 Professional Ethics

Professional ethics consists of a set of norms and values that govern the actions of professionals in an organization. It is based mainly on the universal values that human beings have, applied directly to the work environment and sets guidelines for the performance of the functions of a position within an ethical framework.

Professional ethics is reflected in deontological codes or professional codes through a series of principles and values contained in the form of a Decalogue or documents normally issued by recognized associations in the field of a discipline.

The World Association of Engineers dedicated to standardization and development in technical areas, IEEE (Institute of Electrical and Electronic Engineering), states in

Fig. 48.1 Methodology used for research



its Code of Ethics the importance of technologies and that they affect the quality of life throughout the world and They commit the professionals who make it up to the highest ethical and professional conduct according to ten rules. In particular, the former states, “Maintain the importance of public safety, health, and well-being, strive to comply with ethical design and sustainable development practices, and quickly disclose factors that could endanger the public or to the environment” [5].

The Educational and Scientific Information Society Association for Computing Machinery (ACM) declares that its Code of Ethics is designed to inspire and guide the ethical conduct of all IT professionals and anyone who uses information technology to make an impact. It breaks it down into: general ethical principles, professional responsibilities, professional leadership principles, and code compliance. It is worth highlighting one of its points, which indicates “to carry out comprehensive and exhaustive evaluations of computer systems and their impacts, including an analysis of possible risks” [6].

The Code of Ethics and Professional Practice of Software Engineering was designed as a standard for teaching and practicing software engineering, and provides the ethical and professional obligations of software engineers. It was adopted in 2000 by the IEEE Computer Society and ACM. The code is intended as a guide for members of the profession, committing them to make the analysis, specification, design, implementation, testing and maintenance of software a respected and beneficial profession [7], subject to the eight principles related to conduct and the decisions made by professional software engineers, namely society, customer, product, judgment, administration, profession, colleagues, and staff. In particular, it is worth mentioning the point that engineers should “approve software only if they have a well-founded belief that it is safe, meets specifications, passes appropriate tests, and does not reduce quality of life, privacy, or harm environment. The ultimate effect of work will be the social good” [8].

Ethics is a philosophical discipline that studies good and evil and their relationships with morals and human behavior, gives standards for life, guides behavior and guides decisions. Professional ethics govern conduct and personal activity at the service of society and for the purpose of the common good.

48.2.2 Value Sensitive Design

Value Sensitive Design (VSD) refers to an approach to technology design that takes human values into account in a systematic and principled way throughout the design process [4]. The central focus of focus is on stakeholder analysis, direct and indirect; distinctions between designer values, values

explicitly backed by technology, and stakeholder values; levels of individual, group and social analysis; comprehensive, iterative conceptual, technical and empirical research; and a commitment to progress.

It uses an integrative, tripartite and iterative methodology, which consists of conceptual, empirical and technical research. VSD is primarily concerned with values that focus on human well-being, human dignity, justice, well-being, and human rights.

Conceptual research aims to understand and articulate the various stakeholders of technology, as well as their values and any conflict of values that may arise for these stakeholders through the use of technology. Empirical investigations are qualitative or quantitative design research studies used to inform designers about the values, needs, and practices of users. Technical investigations may include analysis of how people use related technologies or systems design to support the values identified in conceptual and empirical investigations [15].

The approach uses tools like Display Cards [14] that reflect and synthesize a variety of well-established methods in value-sensitive design. With these, the designer is intended to broaden the spectrum of evaluation regarding what he or she should consider as impact elements of the technological project.

Forecast cards are based on a set of four forecast criteria intended to raise awareness of long-term systemic problems in design: Stakeholders, Time, Values, Pervasiveness [14].

- **Stakeholders.** Emphasizes the range of effects of a technology, both on those who are in direct contact with a technology (direct stakeholders), and on those who might not be direct users, but whose lives are nevertheless affected by various interactions around the technology (indirect stakeholders).
- **Time.** Inspired by the long-term perspective of urban planning, the Time criterion helps guide designers to consider the longer term implications of their work – implications that will only emerge after the technology has moved through initial phases of novelty to later phases of appropriation and integration into society.
- **Values.** Emphasizes the impact of technology on human values. Our use of the term values draws from the Value Sensitive Design literature, “what a person or group of people consider important in life.” In interaction design, we have found values of interest to include but are not limited to: autonomy, community, cooperation, democratization, environmental sustainability, expression, fairness, human dignity, inclusivity and exclusivity, informed consent, justice, ownership, privacy, self-efficacy, security, trust, and universal access.
- **Pervasiveness.** Emphasizes systemic interactions that follow from the widespread adoption of an interactive tech-

nology. Technologies can become pervasive with respect to geographic (e.g., city navigation software use within urban areas), cultural (e.g., text messaging within the deaf community), demographic (e.g., online social networking sites among teenagers), and other factors.

Software designers have always considered a technical and economic evaluation when presenting a solution to a particular problem, but it must be recognized that the focus is on the stakeholder “directly” and not necessarily on those who might be affected by it in an indirect way. These considerations and others such as the long-term impact and values such as inclusiveness, equity, human dignity, etc., are necessary to generate a more open, useful and favorable technology for people’s quality of life.

48.3 Results

48.3.1 Ethics as a Subject in an Undergraduate Program

In 1991, a working group combining members of the Computer Association (ACM) and the Institute of Electrical and Electronic Engineers (IEEE) created the 1991 Computer Curriculum (CC91) [9], a framework for a new plan of computer studies. This recommended curriculum contained several units of knowledge related to social and ethical issues in computing. In response to this curriculum, the National Science Foundation (NSF) funded ImpactCS in 1994 [10]. The ImpactCS project brought together a panel of 25 experts in the area to define the core content and tools to integrate social and ethical issues into the computer science curriculum.

The ImpactCS project produced three reports, each of which further defined the roles of ethics and social issues in the new computer science curriculum. The first report adds a tenth subject area to the nine defined by CC91. The second report defines a set of five knowledge units included in the curricular guidelines (responsibility of the IT professional, basic elements of ethical analysis, basic skills of ethical analysis, basic elements of social analysis and basic skills of social analysis) and the third report, provides a pedagogical justification and models for the integration of the material with the existing study plan [10].

In consideration of the importance of ethics in the training of engineers in general, and in the area of computing and informatics in particular, a search was carried out in Chilean universities, which evidenced the incorporation of an ethics or similar subject in their plan of studies.

Sixteen universities were selected, between public and private (see Fig. 48.2), from different regions of the country (see Fig. 48.3). The revised study plans corresponded to 18 engineering majors in the area of computing and informatics,

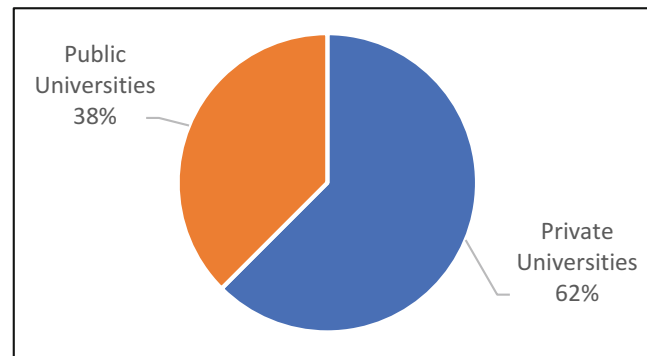


Fig. 48.2 Universities according to public or private administration

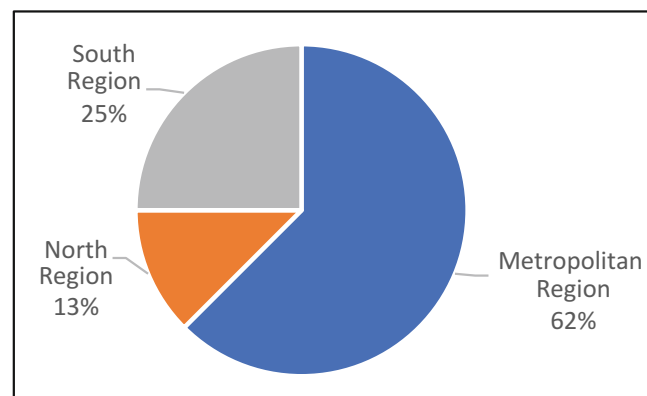


Fig. 48.3 Universities according to geographical location in Chile

but their names are different, as well as their duration, ranging from 8 to 12 semesters.

It is important to note that all the study plans register subjects of Programming, Database, Software Engineering, Engineering Projects, among others. Regarding the Ethics course, only 17% of the courses include a course on this topic (see Fig. 48.4). In two majors, the name of the subject is Professional Ethics, which implies that it is related to the exercise of the profession and the conduct that must govern for the professional who is about to graduate, since, in both cases, it is located in the last or penultimate semester of the degree. Only one of the cases presents the Ethics subject in general terms and is located in the fourth semester of it.

These results show a low incorporation of the topic Ethics in the curricula of the careers reviewed, however it is not possible to deduce that the topic is or is not included in the plan or program through another subject, as its own content and associated with the theme of the subject.

48.3.2 Ethics as a Teaching Methodology

In order to gather more information on the current state of teaching activities related to the incorporation of ethics and

human values in the design of software in careers and/or programs in the areas of computing, informatics, systems or similar, it was carried out a survey that considered academics at the Latin American level, in which 57 professionals from Higher Education Institutions from various countries (see Fig. 48.5).

The basic program considered in the study is Computer Engineering and related careers, that is because at the Latin American level there is a varied range of names for related academic programs, understanding that all of them are considered software design and development subjects, which otherwise they also have various names. As it can be seen in the graph (see Fig. 48.6), the names vary between countries and even within the same country. The programs considered are both undergraduate and graduate, as it is estimated for the research that it is necessary to evaluate the consideration of ethics and human values beyond the training level (see Fig. 48.6).

Regarding the query made among the academic about whether the subject related to software design includes ethics as part of its design, the positive response obtained is 54.4%, compared to 45.6% obtained regarding not incorporated ethics (see Fig. 48.7). It is also important to know if this

topic is included as part of a methodology in the subject, which yielded a result of 21.1%, which reveals that 78.9% do not consider any methodology on the subject (see Fig. 48.8).

These results show low use of a teaching-learning methodology that considers Ethics in subjects related to software design in the curricula of computer engineering or related careers. However, in the open question left to the academics in the survey, to rescue other contributions that the academics could express, all of the respondents agreed that ethics should be included as part of the training of engineers and in some cases, specifically as part of a methodology in software design.

48.4 Analysis and Interpretation of Data

With the arrival of global connectivity and the significant explosion of technological solutions that accompany the daily life of people and society, numerous ethical and social questions arise caused by technology. We live in a world with complex computerized systems, intelligent systems that de-

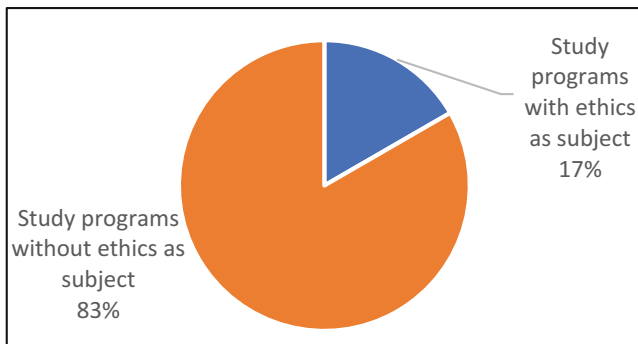


Fig. 48.4 Ethics as a subject in Engineering of Computer and Informatics

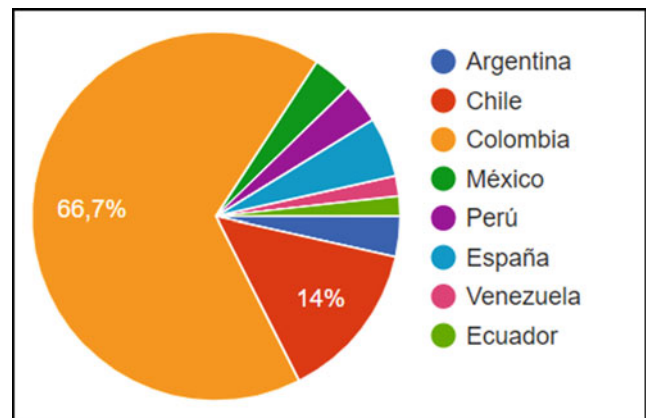
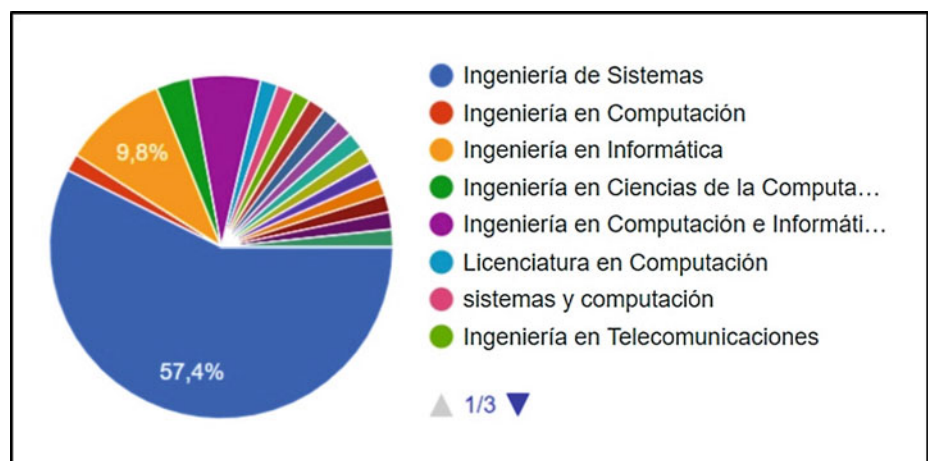


Fig. 48.5 Countries of participating academics

Fig. 48.6 Related academic programs



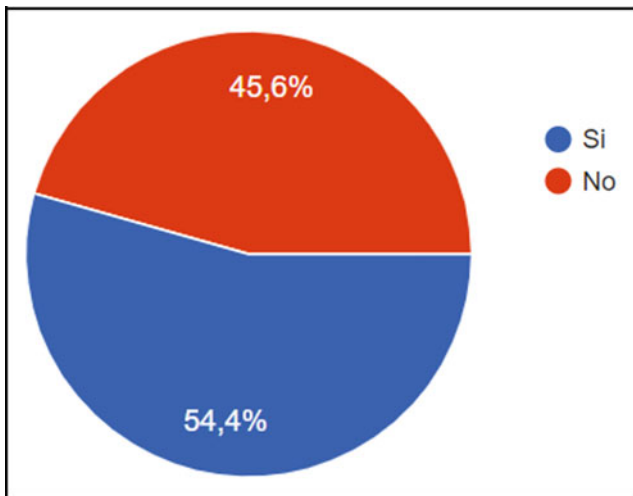


Fig. 48.7 Do you include ethics as part of software design?

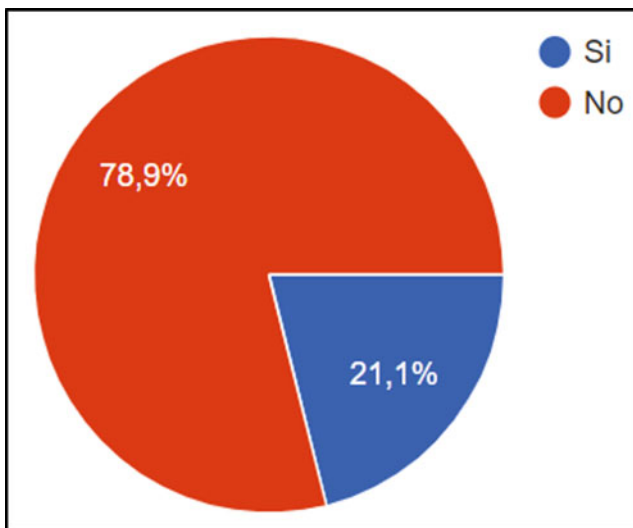


Fig. 48.8 Do you use any methodology to incorporate ethics or human values in software design?

cide for us in some routine activities, such as what is missing in the refrigerator and what to buy, what movie or video to watch, what music to listen to, or other more complex ones such as which fertilizer to use, and in what quantity in a wheat plantation, or what sentence to apply given the crime committed or who to save in the face of an imminent crash in an automated vehicle. Let us recognize, then, that there are numerous ethical and social problems, even in different categories, associated with computer technology and this is visualized with an overwhelming growth.

Forester and Morrison [11] identified the following main categories (or groupings of topics) in which ethical and social concerns generally arise: cybercrime and computer security, theft of software and intellectual property rights, hacking and virus creation, glitch of the computer and the information system, invasion of privacy, social implications of artificial

intelligence and expert systems and computerization of the workplace.

The globalization and inherent nature of these technologies transcend physical and cultural boundaries and, therefore, make it increasingly difficult to comply with accepted laws, regulations and codes of conduct [13]. It is not enough to teach engineering students the globally recognized and accepted codes of ethics; Some methodology that systematically and measurably presents the evaluation of the proposed solution for a certain technological project must be included, using not only technical and economic feasibility, but also including an ethical assessment.

It is relevant that the development of a new technology is defined from its conception and design with a consideration of the personal and social impact seen through human values and ethics. It is necessary to sensitize students, future developers of computer technology, so that they become aware that their decisions regarding what will or will not do a technological solution must be ethical, moral, good and beneficial to society in general [12].

In the particular review of careers or academic programs associated with Computer Engineering or similar in Chile, it is possible to distinguish in a small number of study programs, at least one subject related to ethics, the one that is generally associated or referred to professional ethical codes.

In the Latin American review, academics dedicated to the teaching-learning of software design and development in higher education consider it relevant to include ethics and human values in the conception, design and development of software, which is supported by a little more than half of them. However, most of them do not use a methodology that allows an evaluation of human and ethical values to be applied in the conception, feasibility and design of the technological project.

Engineering students need to develop technical skills related to their discipline, but they also need to develop the ability to recognize and evaluate the social impact that their project will generate, understand the associated cultural, social, legal and ethical problems.

This makes it necessary to include a methodology that contemplates human and ethical values throughout the development process of the new technology, from its conception and design stage. The ImpactCS Project [10] identified two main strategies for teaching computer ethics. These are the integration of the five knowledge units in the existing material at each annual level, and an independent course dedicated to the subject.

The assessment of a methodology based on the Value Sensitive Design approach is visualized as a practical proposal in the training of engineering students in the area of computing and informatics, in the incorporation of evaluations focused on human values, and ethics in design of a technological project.

48.5 Conclusion

There are sufficient arguments to consider the teaching and practice of ethics in the education of engineering students in the area of computing and informatics, not as a Decalogue, but in understanding the impact that a technological solution can affect a person, the community and the society.

Engineering students require developing technical skills related to their discipline, but they also require developing the ability to recognize and evaluate the social impact that their project will generate, understand the associated cultural, social, legal and ethical problems.

Currently, degrees in Computer Engineering and Informatics or similar, most of them do not include subjects related to ethics. The courses that mention a subject with this topic in their study plan refer to professional ethical codes.

Academics dedicated to the teaching-learning of software design and development in higher education consider it relevant to include ethics as part of the subject. However, a high percentage of them do not use a methodology that allows an evaluation of human and ethical values to be applied in the conception, feasibility and design of the technological project.

Given the exponential growth of technology and its presence in people's daily lives, it is urgent to include ethical evaluations when conceiving and designing projects. The human dimension of a product must be part of any software development methodology. It should allow students to verify stakeholders directly and indirectly in it, and to evaluate in an integrated way a solution to the problem under study that is technically and ethically correct.

Value Sensitive Design is presented as a practical theoretical approach to include an assessment focused on the values and the impact that the development of a new technology or technological project will have on stakeholders directly or indirectly.

As future work, it is proposed to generate teaching methodology for the design and development of software in the training of engineering students in the area of computing and informatics.

References

1. Deloitte. Industry 4.0: Challenges and solutions for the digital transformation and use of exponential technologies report, <https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/manufacturing/ch-en-manufacturing-industry-4-0-24102014.pdf>. Last accessed 20 June 2020
2. J. Echeverría, De la filosofía de la ciencia a la filosofía de las tecnologías e innovaciones. *Rev. CTS* **10**(28), 109–119 (2015)
3. Ministry of Science, Technology, Knowledge and Innovation. Government of Chile. Mission and vision, <http://www.minciencia.gob.cl/>. Last accessed 30 June 2020
4. B. Friedman, D. Hendry, *Value Sensitive Design. Shaping Technology with Moral Imagination* (The MIT Press, Cambridge, MA, 2019)
5. IEEE. Code of ethics, <https://www.ieee.org/about/corporate/governance/p7-8.html>. Last accessed 21 May 2020
6. ACM. Code of ethics and professional conduct, <https://www.acm.org/code-of-ethics>. Last accessed 21 May 2020
7. D. Gotterbarn et al., Software engineering code of ethics and professional practice. *Sci. Eng. Ethics* **7**, 231–238 (2001). <https://doi.org/10.1007/s11948-001-0044-4>
8. ACM. The software engineering code of ethics and professional practice, <https://ethics.acm.org/code-of-ethics/software-engineering-code/>. Last accessed 21 May 2020
9. ACM/IEEE-CS Joint Curriculum Task Force, *Computing Curricula 1991* (ACM Press, New York, 1991)
10. C.W. Huff, C.D. Martin, Project ImpactCS Steering Committee, "Computing consequences: A framework for teaching ethical computing" (1st report of the ImpactCS Steering Committee). *Commun. ACM* **38**, 75–84 (1995)
11. T. Forester, P. Morrison, *Computer Ethics: Cautionary Tales and Ethical Dilemmas in Computing*, 2nd edn. (MIT Press, Cambridge, MA, 1994)
12. A. Barnard, C. De Ridder, L. Pretorius, Teaching computer ethics as part of Computer Science and Information Systems curricula (2001)
13. C.D. Martin, E. Yale-Weltz, From awareness to action: Integrating ethics and social responsibility into the computer science curriculum. *Comput. Soc.* **29**(2), 6–13 (1999)
14. B. Friedman, L. Nathan, S. Kane, J. Lin., https://www.envisioningcards.com/?page_id=2#2. Last accessed 29 May 2020
15. B. Friedman, P.H. Kahn Jr., A. Borning, P.H. Kahn, Value sensitive design and information systems, in *Human-Computer Interaction and Management Information Systems: Foundations*, (ME Sharpe, New York, 2006), pp. 348–372

Part IX

Pandemic

Using UAV, IoMT and AI for Monitoring and Supplying of COVID-19 Patients

A. J. Dantas, L. D. Jesus, A. C. B. Ramos, P. Hokama, F. Mora-Camino, R. Katarya, O. P. Verma, P. K. Gupta, G. Singh, and K. Ouahada

Abstract

The Corona Virus Disease 2019 (COVID-19) is an infectious disease caused by a newly discovered corona virus SARS-CoV-2. It is similar to the flu and raises concerns about the alarming levels of spread and severity, resulting in a continuous pandemic worldwide. In eight months, it infected 90 million people worldwide and more than 2 million died. Unmanned Aerial Vehicles (UAVs) can be very useful in supporting logistical support for the COVID-19 pandemic. This work aims to investigate UAV-based applications for logistical support to situations caused by the COVID-19 pandemic and proposes an architecture to deal with pandemic situations in different scenarios using real-time case studies.

Keywords

UAV applications · Spread monitoring · Internet of medical things and machine learning

A. J. Dantas (✉) · L. D. Jesus (✉) · A. C. B. Ramos · P. Hokama
Federal University of Itajubá, Itajubá, Brazil
e-mail: antoniodantas@unifei.edu.br; d2019102840@unifei.edu.br;
ramos@unifei.edu.br; hokama@unifei.edu.br

F. Mora-Camino
Federal Fluminense University, Niterói, Brazil
e-mail: felixmora@id.uff.br

R. Katarya · O. P. Verma
Delhi Technological University, Delhi, India
e-mail: rahulkatarya@dtu.ac.in; opverma.dce@gmail.com

P. K. Gupta
Jaypee Information Technology University, Noida, India
e-mail: pradeepkumar.gupta@juit.ac.in

G. Singh · K. Ouahada
University of Johannesburg, Johannesburg, South Africa
e-mail: ghanshyams@uj.ac.za

49.1 Introduction

This paper presents an International Cooperation Project that aims joining efforts of research groups of BRICS countries (Brazil, Russia, India, China and South Africa) to use Unmanned Aerial Vehicles—UAV, equipped with electronic systems to ensure, precisely, its position and trajectory control to collect images to be transmitted to a control station at a distance for cloud storage and subsequent digital processing.

Current UAV-based systems have proven to be of great use for monitoring people due to their mobility, low cost and integration of electronic systems that can expand the capabilities of measuring social distance, COVID-19 monitoring and data collection using Artificial Intelligence (AI), thermal image, cleaning with data analysis, record keeping etc. Therefore, the general objective of the project is to develop methods, techniques and tools supported by intelligent algorithms, which will help to combat the COVID-19 pandemic. Therefore, it will be necessary to understand the need and the requirements for improvements in UAV-based systems for the development of a system for intelligent health care, the main objectives of this work are:

1. Propose noise reduction strategies that allow the movement of simple UAVs in trajectories (in “indoor” and “outdoor” environments) based on collision-free zones.
2. Collaborate in strategies to inform the population about the health risks presented by COVID-19.
3. Propose a system based on artificial intelligence that collects data through UAVs, analyzes and provides the necessary security measures.
4. Propose a multilayer architecture that collects UAV information and allows for the sharing of data and necessary analysis.

5. Simulate the UAV-based system for COVID-19 operations, such as monitoring, control, thermal imaging, sanitation, social distance, medication, data analysis and generation of statistics for the control room.
6. Implement a system based on UAVs in real time for hygiene, monitoring, surveillance, facial recognition, temperature measurement in COVID-19 combat environments.
7. Design and display UAV-based smart health system statistics in a local monitoring and control room.

49.2 Relevance and Impact

In December 2019, a new Coronavirus Disease (COVID-19) erupted in Wuhan (Hubei, China) and spread rapidly from a single city to the entire country. It did not take long for this epidemic to spread throughout the world. After that, the World Health Organization (WHO) declared this epidemic disease to be a pandemic.

COVID-19 is a highly contagious infectious respiratory disease that can cause respiratory, physical and psychological dysfunction in patients. Respiratory rehabilitation reduces the patient's dyspnea symptoms, relieves anxiety and depression, reduces the need for hospital application, increases functional capacity and improves the patient's quality of life. SARS-CoV-2 transmission occurs frequently in hospital settings, with several cases of hospital transmission reporting highlighting the vulnerability of healthcare professionals.

The unexpected appearance of COVID-19 has been forcing governments to make changes in parameters both in the approach and in the solution of several problems inherent to the treatment of the pandemic. In this context, the use of UAVs for social health can be of fundamental importance, especially for coping with the COVID-19 epidemic in the world.

Humanity is living in unprecedented times when almost the whole world was affected by COVID-19. Around the world nurses, doctors and health professionals in general are working hard to help diagnose patients, several government leaders have suggested maintaining social distance in order to avoid the explosion of cases of contamination. The police and health units have been conducting aerial inspections trying to raise public awareness, along with many other measures to be taken at all levels [1, 2].

Drones are proving to be of great help at different levels. Several countries are considering UAVs to be of great use through various measures [1–5]. Cradle of Pandemia and one of the most developed countries in terms of Drones, several applications have been developed in China, for example, drones have been used for disinfection by spraying disinfectant on public roads where, in addition to hygiene, monitoring

is also carried out specific areas. In the surveillance area where more than 100 drones are used in cities to prevent viral transmission by carriers, especially if the distance between individuals becomes less than a specific value (for example 1.5 m) or if people are walking in locations masked audiences (similar practices were also followed in Spain, Kuwait and the United Arab Emirates) [6].

In the United States, the current epicenter of the pandemic, a personal medical kit for COVID-19 is being delivered by drones to remote locations [6]. Like other drone-based systems, this system is effective in providing medical services and other supplies. In the USA, the use of drones has been shown to be effective in rural areas, where COVID-19 symptoms are found in patients.

In Brazil, despite its position as second in both the number of infected and the number of deaths (second only to the United States), the main use of UAVs has been carried out by police agencies in the monitoring of public places in order to avoid crowds of people. The use of UAVs for other possible applications, comes up against the severe restrictions required by the governmental aeronautical authority, which makes it difficult to diversify applications. Currently, only one company has been testing the use of drones to combat the pandemic, using drones to distribute medical supplies in places that are difficult to access.

At the forefront of countries that use drones to support services demanded by COVID-19, India has already been conducting tests in states like Delhi, Kerala and Assam that are making announcements during city surveillance via drones. In the state of Maharashtra, it is a step ahead, as data analysis reports are generated from the area covered by drones [7, 8] in general observations, government officials in India have given special permissions to their civil servants and police to use drone-based technology for surveillance, monitoring, medication, hygiene, data analysis, reporting and future decisions.

Australia has been efficiently controlling the pandemic, tests have been carried out with drones where, during a flight, through the city it is possible to detect whether someone has a "questionable" breathing pattern or not [3]. Sensors installed in drones also record body temperature, heart rate, breathing and other abnormalities. These measurements are carried out in different areas, especially in overcrowded areas.

Therefore, from the actions described above, the importance of using UAVs is evidenced both in the cleaning of environments, patient support and screening and in the distribution of medical supplies in places of difficult access. It is hoped that with the development of this project, it will be possible to integrate the various potentialities into an integrated system for the administration and control of medical supplies and patient monitoring, both in closed and open environments.

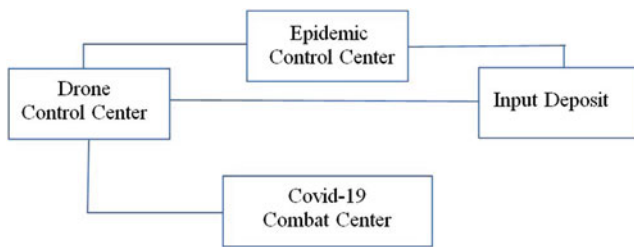


Fig. 49.1 System architecture. Source: Author

49.3 Methodology

This project is part of a larger project carried out by a group of Brazilian, Indian and South African professors/researchers that aim to develop strategies to combat COVID-19. Professors and researchers from the three countries should interact from technical visits (30 days) and evaluation meetings (7–10 days) aiming at sharing and exchanging information in order to ensure that everyone has mastery of the methods, techniques and tools developed by the different work teams.

The Brazilian team proposes the development of an architecture based on UAVs to support logistics, using integrated AI tools for monitoring, control, delivery of inputs, sprinkling environments etc., see Fig 49.1. For this purpose, machine learning and deep learning techniques will be applied to define UAV routes, identify objects such as: landing points, crowds of people, hygiene etc. From the analysis of image data in real time combined with Information Systems for decision making.

AI is applied to the identification of contaminated patients, logistics for the distribution of inputs and logistics. Subsequently, IS approaches are applied to efficiently store and process data from locations on the pandemic front to the control center located far from the COVID-19 hotspots, even more within the municipality. Likewise, other technological aspects, including monitoring of the movement of people, profile of contaminated etc. are recovered, observed and analyzed for the spread of the pandemic.

The Epidemic Control Center, performs the control and monitoring of the epidemic from data collected in real time by the UAVs during the flyover missions of areas pre-defined by the flight mission controllers defined by requests received from the Control Center. The data collected by the UAVs is transmitted to the Epidemic Control Center for evaluation, which can result in the following tasks:

1. Replacement of hospital supplies, in this case, depending on the need and urgency, the supplies can be sent directly by the UAVs to the Combat Centers. In this case, UAVs must be sent to the Input Deposit for loading;
2. Isolation of people, according to body temperature data provided by thermal cameras from aerial monitoring in

areas predefined by the Control Center (external access ordinances, waiting lines, etc.);

3. Sprinkling of chemical sanitary agents in environments with a large number of people at any given time;
4. Identification of people who are not complying with the rules of social distance and/or move without wearing masks.
5. Dispersion of people during agglomerations.

In the context of action with the population, in order to guide them about the risks of the pandemic and preventive actions, it will be necessary to identify health policies and community agents or community leaders, future multipliers of knowledge about prevention strategies against COVID-19.

Once identified, they will be guided on the methodology and applicability of the material to be delivered via UAVs (outdoor), as well as the definition of the amount of material to be distributed. The distribution of this material will take place on a daily, weekly or fortnightly basis according to the established schedule. The materials will come from agencies, trustworthy, governmental and non-governmental and will consist of: information sheets, videos, booklets, educational programs etc.

49.4 Partnerships and Collaborations Already Established

Since 2008, VISCAP professors and researchers have been working together with Dr. Félix Mora-Camino, formerly professor at the Ecole Nationale de l'Aviation Civile - ENAC and since 2020 professor at the Universidade Federal Fluminense, always collaborating in the areas of UAV trajectory control systems. In two opportunities (2008 and 2012) Félix collaborated with UNIFEI professors and researchers by teaching two classroom courses on Aircraft Control Systems. Her interaction with UNIFEI professors and researchers has already produced several joint publications with UNIFEI both in international journals and in national and international conferences.

As of 2015, a collaboration between VISCAP professors and researchers began with professors and researchers from Jaypee University of Information Technology—JUIT led by Dr. Pradeep Kumar Gupta, especially in applications of Internet of Things (IoT) and Cloud Computing (Cloud Computing) aiming to implement algorithms to identify objects in large quantities of photographs obtained from cameras embedded in UAVs as a result of the interaction of Gupta, two articles have been published at international conferences. In 2017 Gupta offered a face-to-face course for UNIFEI professors and researchers on the development of Cloud Computing algorithms for classifying and identifying objects in images.

The interaction with professors and researchers in the medical field at the University of Vale do Sapucaí—UNIVAS began in 2017 with the development of a Computational System for Assessing Fragility in the Elderly, resulting in two Master's dissertations (one in each university) in addition articles published at conferences and a national patent for the developed system.

49.5 Expected Results

The products expected as a result of this project until the end of the first year are:

- (a) The development of noise reduction strategies caused by UAV propellers, especially in closed environments.
- (b) The development of a software and hardware system to control aircraft trajectories for the delivery of inputs, cleaning environments and monitoring the behavior of individuals both in closed and open environments.
- (c) The development of a software system for processing multispectral images referring to the body temperature of human beings;
- (d) The implementation of specific Flight Plans for monitoring possible patients with COVID-19 for the city of Itajubá and region;

At the end of the second year of the project, it is expected:


- (a) Training of personnel for the operation of UAVs and the Control Center;
- (b) The development of efficient strategies for the logistics of distribution of hospital supplies;
- (c) The development of procedures for sprinkling products to clean environments;
- (d) The identification of COVID-19 patients who do not comply with social isolation.

49.6 Conclusion

This work in progress will allow cooperation between the countries Brazil, India and South Africa, in an exchange of expertise, whose own technologies are implemented, thus contributing to global science in the fight against COVID-19, using UAVs as a support.

References

1. A. Kumar, K. Sharma et al., A drone-based networked system and methods for combating coronavirus disease (COVID-19) pandemic. Cornell University. arXiv: 2006.06943 submitted on 12 Jun 2020. <https://arxiv.org/abs/2006.06943>
2. S. Tuli, S. Tuli, R. Tuli, S.S. Gill, Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things* **11**, 100222 (2020). Elsevier. <https://doi.org/10.1016/j.iot.2020.100222>
3. Draganfly selected to globally integrate breakthrough health diagnosis technology immediately onto autonomous camera's and specialized ARPs to combat coronavirus (COVID19) and future health emergencies. [Online]. Disponível em: <https://apnews.com/Globe%20Newswire/dc01344350423d7d64c99ebbe8fb7548>
4. ARPs and the coronavirus: How ARPs are helping containment efforts [Online]. Disponível em: <https://uavcoach.com/ARPs-coronavirus/>
5. Corona combat ARP: Indian robotics solution launches corona combat ARP to fight covid-19, government news, et government [Online]. Disponível em: <https://government.economictimes.indiatimes.com/news/technology/indianrobotics-olution-launches-corona-combat-ARP-to-fightcovid-19/75077517>
6. ARP technology: a new ally in the fight against COVID-19 [Online]. Disponível em: <https://www.mdlinx.com/internal-medicine/article/6767>
7. ARP Association of Kerala: Últimas Notícias e Vídeos, Fotos sobre ARP Association of Kerala | O econômico Times. <https://economictimes.indiatimes.com/topic/ARP-Association-of-Kerala> (acessado em 15 de abril de 2020)
8. Bloqueio da Covid-19: as autoridades confiam no ARP eye para manter a vigília - The Economic Times. <https://economictimes.indiatimes.com/news/politics-and-nation/covid-19-lockdown-authorities-rely-on-ARP-eye-to-keep-vigil> (acessado em 15 de abril de 2020)

Debdeep Dey , Sarangi Patel, Karasani Tharun Kumar Reddy, and Siddharth Sen

Abstract

In this paper, we conduct mathematical and statistical analysis to answer the following questions for COVID-19 : (Q1) How has the pandemic progressed since March 2020? (Q2) How effective have the social distancing approaches been? (Q3) How has the pandemic spread differently in different Indian states? (Q4) What would be the impact of a vaccination campaign on the current state of the disease spread? (Q5) What are the economic and ethical nuances that need to be considered when undertaking such a campaign? (Q6) Whom do we prioritize for vaccination? (Q7) How long does the immunity from the vaccine likely to last? (Q8) How important is it to track the mutations of the virus?

Keywords

COVID-19 · Time-dependent SIR · SIRV · HIT · Vaccination strategy · SARS-CoV-2 · Vaccine distribution strategy · Compartmental models

50.1 Introduction

This paper has tried to answer the questions mentioned above analytically, as they are important and relevant in the current socio-economic climate. Through this analysis, we also aim

D. Dey (✉) · S. Sen
Kolkata, West Bengal, India

S. Patel
Ahmedabad, Gujarat, India

K. Tharun Kumar Reddy
Guntur, Andhra Pradesh, India

to resolve some of the personal misconceptions that we, the authors, had about COVID-19. One's ability to answer each of these questions would depend, significantly, on their understanding of the disease, from:

- an epidemiological perspective
- a statistical perspective

Hence, to strengthen our statistical understanding of the disease, we decided to compute its progression in Sects. 50.2.2.2 and 50.5.1 using variants of the popular SIR (Susceptible, Infected, Recovered) model:

- Time-Dependent SIR
- SIRV

Having understood the spread of the disease and the impact of a vaccination campaign on the spread, in Sect. 50.8 the focus shifted to the next logical step:

- Ensuring that the vaccination campaign is carried out in an optimal and ethical manner

And to this end, we have:

- proposed a novel algorithm for ensuring an equitable, fair and effective distribution of the vaccines
- identified nuances—economic and ethical—that should be considered while carrying out the process

such that the benefits of the initiative can be perceived both tangibly and measurably.

50.2 What Is the Current State of the Infection Spread?

50.2.1 The Rationale Behind the Choice of Models

Epidemiology has progressed in leaps and bounds since John Snow's famed investigations into the causes of cholera in the nineteenth century [1] and now forms the basis of modern public health. It has progressed beyond many other fields such that now it has become routine to incorporate mathematical and computational methods into the study of disease processes [2]. This approach is particularly powerful when it comes to pandemics as these outbreaks affect vast numbers of people and can spread rapidly. Outbreaks such as the current one are likely to become more common with climate change, movement of people, and failing antibiotics [3]. Mathematical modeling, which can predict disease progress and outcome as well as identify the potential causes of transmission and optimal interventions, is an increasingly important implementation in the toolkit of modern epidemiology. The outbreak caused by the SARS-COV2 virus needs to be modelled with an appropriate level of abstraction for us to be able to get around the data availability constraints [4] while still being able to understand and predict the spread of the disease and its eventual containment reliably.

50.2.2 Modeling the Spread of the Disease

50.2.2.1 Compartmental Models

Compartmental models are used to simplify the process of mathematical modeling of infectious diseases wherein the population under investigation is divided into compartments or systems; the flow of actors between these systems are modelled using different probabilistic parameters. The rate of population change in each of these systems is determined using Ordinary Differential Equations (ODE) and are deterministic [5] in nature.

50.2.2.2 SIR Model

The SIR model, dating back to the original work of A.G. McKendrick and W.O. Kermack in 1927 [6], comprises of a simple deterministic model predicting the outbreak of epidemics. The model uses the following the system of ODEs:

$$ds/dt = -\beta(s(t)x(t))/n, \quad (50.1)$$

$$dx/dt = \beta(s(t)x(t))/n - \gamma x(t) \quad (50.2)$$

$$dz/dt = \gamma x(t) \quad (50.3)$$

respectively, with the initial conditions

$$s(0) = n_1 \geq 0$$

$$x(0) = n_2 \geq 0$$

$$z(0) = n_3 \geq 0$$

$$n_i \in \mathfrak{R}, i = 1, 2, 3,$$

and where the infection transmission rate β and the mean recovery rate γ are positive constants. This model considers a fixed population divided into three compartments: susceptible(S) $s(t)$, infected(I) $x(t)$, and recovered(R) $z(t)$, respectively. The compartments used can be defined as:

- $s(t)$ represents the number of individuals who have not been yet infected with the disease at time t . These are the individuals who are susceptible to the disease,
- $x(t)$ represents the number of individuals who have already been infected, and are capable of spreading the disease to those in the susceptible category, and
- $z(t)$ comprises of individuals who had been infected but have since then recovered from the disease. It is assumed that those who have recovered from the disease cannot be susceptible again or spread the infection to others.

The initial conditions $s(0) = n_1$, $x(0) = n_2$ and $z(0) = n_3$ must satisfy the condition:

$$n_1 + n_2 + n_3 = n, \quad (50.4)$$

where n is the total fixed population. β and γ are the transition rates between the individual compartments. The transition rate between S (Susceptible) and I (infected) is β , where β is the transmission rate, which denotes the probability of contracting the disease when a susceptible person comes in contact of an infectious subject [7, 8]. The transition rate between I (Infected) and R (recovered), is γ , which denotes the rate of recovery or death. If the duration of the infection is represented as D , then $\gamma = 1/D$, since an individual recovers in D units of time [9, 10].

β and γ are interpreted as transition rates (probabilities) and hence their range is defined by $0 \leq \beta \leq 1$ and $0 \leq \gamma \leq 1$, respectively.

The epidemiological threshold, the key value governing the time evolution of these equations, is defined as, $R_0 = \beta/\gamma$, more commonly referred to as the Reproduction Number.

R_0 can thus be defined as the number of secondary infections caused by a single primary infection; it determines the number of people infected by contact with a single infected person before his death or recovery.

When $R_0 < 1$, each person who contracts the disease will not infect anyone else before dying or recovering, so the outbreak will peter out.

When $R_0 > 1$, each infected person will infect more than one person, so the pandemic will spread.

R_0 is one of the more important quantities in epidemiology.

50.2.2.3 Time Dependent SIR Model

The traditional SIR model has two time-invariant variables, the transmission or contact rate β and the recovery rate γ . The SIR model generally neglects the time-varying property of these variables while one can intuitively understand that these variables are expected to change with time given change in the population's behaviour due to voluntary or enforced social distancing, changing government regulations and load on the healthcare infrastructure. Most data driven methods used to predict the spread of the disease seem to perfectly fit the data, often at the risk of over-fitting to the training data. We wanted to capture the time variant nature of the aforementioned variables and implemented the time-dependent SIR (TSIR) model as seen in literature [11]. Replacing β and γ by $\beta(t)$ and $\gamma(t)$ in the differential equations 50.1, 50.2 and 50.3 yields

$$ds/dt = -\beta(t)s(t)x(t)/n, \quad (50.5)$$

$$dx/dt = \beta(t)s(t)x(t)/n - \gamma(t)x(t), \quad (50.6)$$

$$dz/dt = \gamma(t)x(t), \quad (50.7)$$

The intuition behind these equations remain almost the same as before, $s(t)$ denotes the number of susceptible people in the population at time t . We assume that the total population is n and the probability of a randomly chosen person being susceptible is $s(t)/n$. Hence, any infected individual will, on an average, contact $\beta(t)s(t)/n$ people in the susceptible state per unit time t , which further implies that the number of newly infected people is $\beta(t)s(t)x(t)/n$. The number of people in the susceptible compartment will decrease by the same amount. Thus infected, people will eventually recover and move out of the infected compartment a rate of $\gamma(t)x(t)$. COVID-19 data being updated in days, the differential equations in 50.5, 50.6, 50.7 need to be revised into discrete time difference equations:

$$s(t+1) - s(t) = -\beta(t)s(t)x(t)/n, \quad (50.8)$$

$$x(t+1) - x(t) = \beta(t)s(t)x(t)/n - \gamma(t)x(t), \quad (50.9)$$

$$z(t+1) - z(t) = \gamma(t)x(t) \quad (50.10)$$

The three variables $s(t)$, $x(t)$, $z(t)$ still satisfy Eq. 50.4. Given that most of the population is in the susceptible state at the beginning of the pandemic, it is safe to assume that $s(t) \approx n$, $t \geq 0$, and with this assumption, difference equation 50.9 further simplifies to:

$$x(t+1) - x(t) = \beta x(t) - \gamma x(t), \quad (50.11)$$

And from Eq. 50.11, the value of $\beta(t)$ and $\gamma(t)$ for each day can be derived as follows:

$$\beta = (z(t+1) - z(t))/x(t) \quad (50.12)$$

$$\gamma = ([x(t+1) - x(t)] + [z(t+1) - z(t)])/x(t) \quad (50.13)$$

We can now use historical data from a certain period and machine learning methods to track and predict the time varying transmission and recovering rates $\beta(t)$ and $\gamma(t)$ by the commonly used Finite Impulse Response (FIR) filters in linear systems. The predicted transmission and recovery rates as predicted by the FIR filter can be represented by $\hat{\beta}$ and $\hat{\gamma}$ and are defined as:

$$\begin{aligned} \hat{\beta} &= a_1\beta(t-1) + a_2\beta(t-2) + \dots + a_j\beta(t-j) + a_0 \\ &= \sum_{j=1}^J a_j\beta(t-j) + a_0 \end{aligned} \quad (50.14)$$

$$\begin{aligned} \hat{\gamma} &= b_1\gamma(t-1) + b_2\gamma(t-2) + \dots + b_j\gamma(t-j) + b_0 \\ &= \sum_{k=1}^K b_k\gamma(t-k) + b_0 \end{aligned} \quad (50.15)$$

where J and K are the orders of the two FIR filters ($0 < J, K < T_2$), $a_j, j=0,1,\dots,J$ and $b_k, k=0,1,\dots,K$ are the coefficients of the impulse responses of these two FIR filters. Ridge regression is then used to estimate the coefficients of the impulse response of the FIR filter by solving the following optimization problems:

$$\min a_j \sum_{t=j}^{T-2} (\beta(t) - \hat{\beta})^2 + \alpha_1 \sum_{j=0}^J a_j^2 \quad (50.16)$$

$$\min b_k \sum_{t=K}^{T-2} (\gamma(t) - \hat{\gamma})^2 + \alpha_2 \sum_{k=0}^K b_k^2 \quad (50.17)$$

where α_1 and α_2 are regularization parameters.

50.3 Results and Inferences

50.3.1 Data-Sets and Parameter Setup

Name	Value
Source	https://www.covid19india.org/
Orders beta	30
Orders gamma	30
Ridge regression	Scikit-learn library
Training starting date	3rd March 2020
Test data from	2nd Oct to 22 Oct 2020
Training till	1st October 2020
Prediction till	120 days from 23rd October 2020
Vacc release date assumption	11th November 2020
α_1	0.0003675
α_2	0.0001675
Stop day	120 (stopping criteria for prediction)

50.3.2 What Has Been the Impact of Social Distancing Policies Imposed by the Government?

Analysing the trend of the β and the γ values helped to understand the changing transmission rate of the virus and the recovery rates of those infected as the government continued to implement prolonged lock-down phases [12] in the country and also as these lock-down periods were gradually lifted [13]. The TSIR model helped calculate these values as seen in Fig. 50.1 where L_i and U_i represent the i th period of Lockdown and Unlock respectively. As is visible in the graph, the transmission rate β decreases drastically with the imposition of the initial lock-down phases and then gradually with the subsequent ones. This is expected as the interactions between individual actors reduce because of social distancing [14], and hence the probability of spreading the disease decreases. As these lock-down restrictions were gradually lifted off by the government, the value of β witnessed a sudden increase. The recovery rate, γ , can be observed to have been gradually increasing; this can possibly be explained by the fact that as the knowledge base for treating the disease has improved with time and research and the load on the healthcare infrastructure has reduced with the “flattening the curve” initiatives, patients have received more efficient and effective treatment [15].

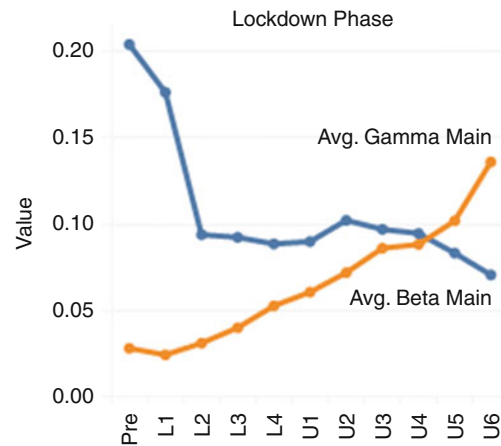


Fig. 50.1 Change in β and γ during and after lockdown in India

50.3.3 Time Evolution of the Disease Spread

The TSIR model was trained and tested on Pan India data as well data for each state in the country. Figure 50.2 visually demonstrates the evolution of the number of infected and recovered cases in the country since March 2020 and starts predicting the same values from November 2020. R_0 is defined as β/γ in the classical SIR model; this is because an infected person takes (on average) $1/\gamma$ days to recover from the disease, and during the same period, he is expected to come in contact with β people on average. In the TSIR model, the basic reproduction number $R_0(t)$, as seen in Fig. 50.3, is a function of time, and it is defined as $\beta(t)/\gamma(t)$.

50.3.4 Accuracy Metrics

To evaluate the accuracy of our model, we compared one-day predictions of infected and recovered people with the actual numbers. The percentage error between these numbers was calculated for each day. The model fit is nearly perfect with about 5% error margin. The percentage error in R_0 does not exceed 1%.

We further examine the precision of our model by using two statistical metrics • Mean Error (%): Average of percentage errors between the predicted and actual values of infected people. Mean Error for our model turns out to be **1.00%** • Mean Squared Error (%): Average of the squares of errors between predicted and actual values. We report a MSE of **2.22%**

Fig. 50.2 TSIR output for India. The dark green line is the actual number of recovered cases and the light green line is the plot of the predicted values. The dotted red line marks the start of the predictions provided as part of the model output. Similarly, the dark red line shows the actual number of infected cases and the lighter red line shows the predicted values

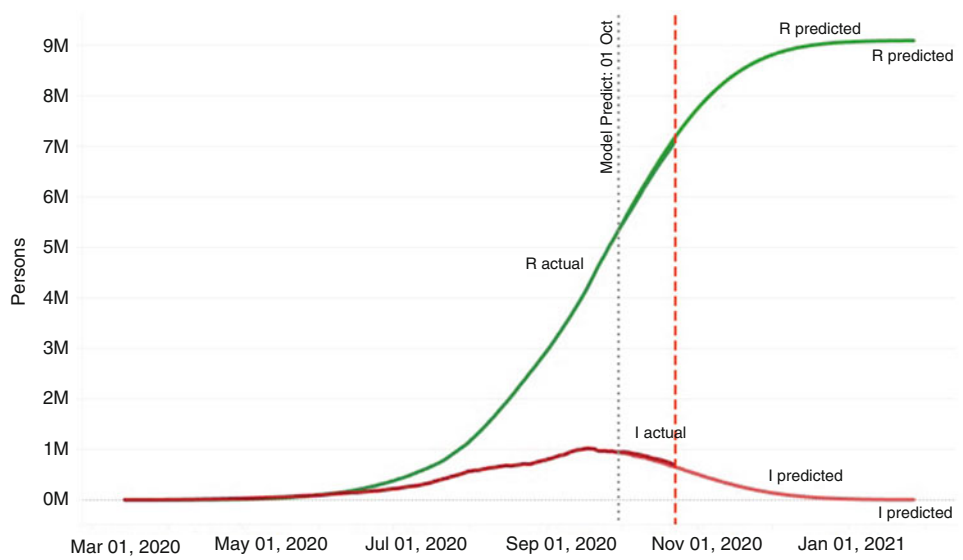
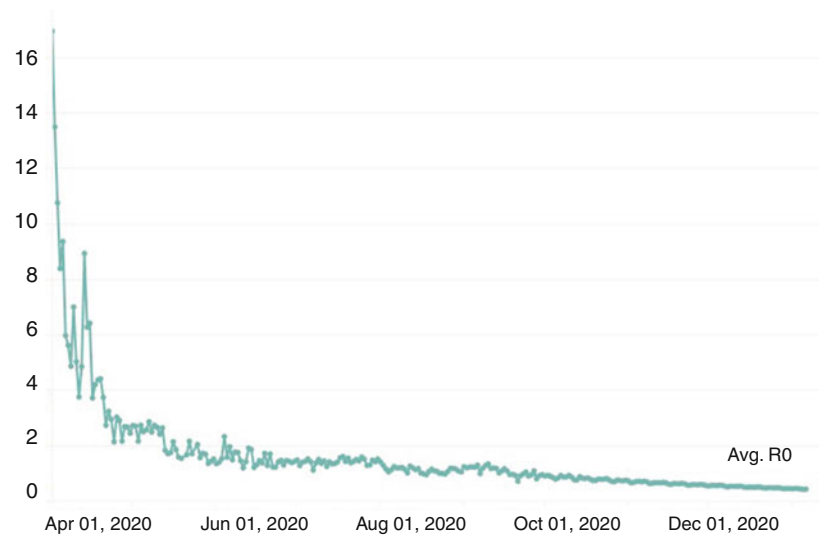


Fig. 50.3 Predicted R_0 for India. The time dependent value of $R_0(t)$ plotted against time. A decreasing trend is being predicted



50.4 Limitations of Deterministic Models

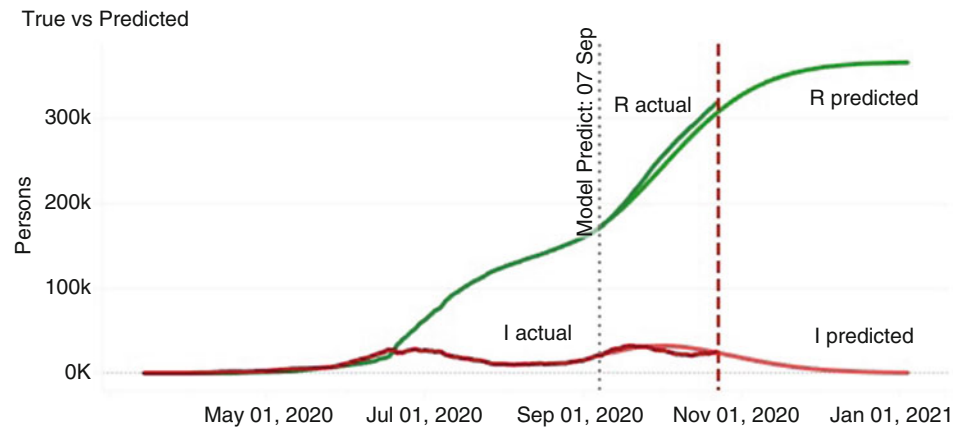
There have been numerous instances of subsequent outbreaks of the pandemic in various regions of India and the world. Our current TSIR model, being a deterministic one, can predict these subsequent outbreaks, but only about 2–3 weeks in advance. This is a well documented limitation of deterministic models and it has been shown that such models can only provide accurate forecast for ≈ 2 weeks from the last training date [16]. Delhi, an UT in India, has had a second outbreak. We decided to train our model till about 3 weeks before the second peak and test if the model correctly predicts this subsequent peak. As seen in Fig. 50.4, the TSIR model can predict the outbreak within a reasonable accuracy threshold. Although there is no evidence that people who have recovered from the disease are protected from re-infection [17], we

have assumed in our model definition that recovered people cannot be susceptible to the disease again; this is primarily because of the unavailability of reliable data-sets that capture this phenomenon of reinfection. We would also like to point out that our epidemiological analysis is constrained by the prevalence or lack thereof of testing in each of these regions [18].

50.5 How Do We Contain COVID-19?

While vaccination is likely to be effective in controlling the spread of the disease, we want to statistically model its impact. The susceptible population when vaccinated would be moved to a new compartment, the vaccinated state, and would thus be immunised against the virus. This would make

Fig. 50.4 Predicted second outbreak in Delhi



people in this compartment equivalent to the population in the recovered bucket, the assumption being, that they would neither be susceptible again nor would spread the infection to other susceptible people [19].

50.5.1 SIRV Model

We build on the SIR model by introducing another compartment V [20] and a vaccination parameter p to model the impact of vaccination using the following ODEs:

$$ds/dt = -\beta(t)(s(t)x(t))/n - ps(t) \quad (50.18)$$

$$dx/dt = \beta(t)(s(t)/n)x(t) - \gamma(t)x(t) - ps(t) \quad (50.19)$$

$$dz/dt = \gamma(t)x(t), \quad (50.20)$$

$$dv/dt = px(t) \quad (50.21)$$

where p is the rate of vaccination. The initial conditions, $s(0) = n_1$, $x(0) = n_2$, $z(0) = n_3$ and $v(0) = n_4$, as seen before, are not independent and must satisfy the following condition:

$$n_1 + n_2 + n_3 + n_4 = n \quad (50.22)$$

where n is the fixed population. All other parameter and variable definitions remain unchanged.

50.6 What Would Be the Impact of Vaccination on the Disease in Its Current State?

To simulate the impact of vaccination on the current state of the disease, we use the predictions from the TSIR model to estimate the number of infections $x(t)$ in different states of India; for each state the Herd Immunity Threshold (HIT) [21] is calculated. The HIT can be used to estimate the number of vaccines required for a particular state (HIT*Population of the state) and the SIRV model is finally leveraged to simulate

the impact of vaccination at a predetermined rate p as mentioned in the differential equation 50.21. Figure 50.5 demonstrates the potential impact of vaccination on the number of infections. At the time of writing this paper, Karnataka has a population of ≈ 67 Million, R_0 of 1.1 and a HIT of 10.59%. Even at a vaccination rate of 1%, the number of infected cases would reduce significantly in a relatively small amount of time, assuming sufficient vaccine availability. Delhi, on the other hand, has a R_0 of 0.8 and hence the spread of the virus is likely to be organically contained. Hence Delhi may not be prioritised for vaccination.

50.7 The Economics of Vaccination in India

India, the second most populous country in the world [22], has the second highest number of COVID-19 infections recorded worldwide at 7.9 Million cases at the time of writing this paper. Clearly the country, like many other developing economies, is in dire need of vaccines to curb the infection spread while it tries to navigate its way to a relatively stable economy, one that had to take a substantial hit [23] owing to lock-downs and other regulations which were imposed to curtail the infection spread and reduce the shock on the existing healthcare infrastructure in the country. Having said that, the country is at the forefront of the of the process to develop and procure vaccines for its population. All over the world, there are more than 140 COVID-19 candidate vaccines available under various stages of development [24]. The Drug Controller General of India has recently approved 2 vaccines for restricted emergency use in the country.

50.8 Whom Do We Prioritise for Vaccination?

Any successful vaccination campaign, in our opinion, should be able to ensure an equitable, transparent and effective delivery of the vaccine. As of now, the government bodies have

Fig. 50.5 SIRV Output for the state of Karnataka. The lighter red line shows the expected decrease in the number of infected people as a result of vaccinating; here the rate of vaccination is 1%

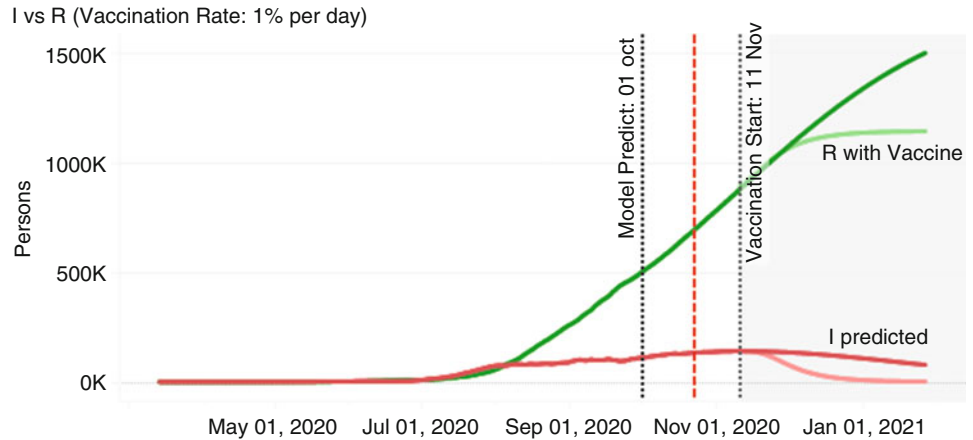
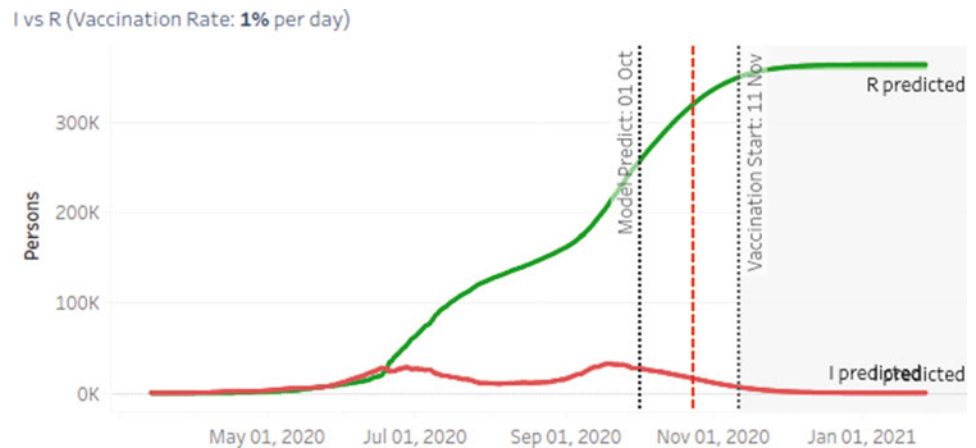


Fig. 50.6 SIRV Output for the Union Territory of Delhi



mentioned that when the time comes for such a vaccination campaign, it would be carried out by the Central Government bodies and the individual states would not be asked to chart their own pathways for procurement or distribution of the vaccine. Ensuring the last mile delivery of the vaccines would be imperative for the success of such a campaign given the poor reach of healthcare infrastructure in India's remote areas. Developing a digital infrastructure for managing vaccine inventory and delivery, including tracking the vaccination process real-time would also be crucial. Aadhar ids, (~ 89% of the population has them), could be leveraged for tracking vaccinations at an individual level.

When it comes to prioritising certain demographic buckets for vaccination, we have identified the following three in decreasing order of priority:

- People with higher number of social interactions (health-care workers/police officers)
- People in a vulnerable age bucket (elderly and those with co-morbidities)
- People in a vulnerable economic bucket (the poor and under-nourished with low immunity)

50.8.1 Proposed Algorithm for Vaccine Distribution

In this paper, the following novel algorithm is being proposed that can be used to carry out vaccination campaigns in an ethical and optimal manner:

1. Step 0: Identify problematic states from the TSIR, SIRV models and correlate findings with available testing data to ensure that the spread of infections in different states is captured in an unbiased manner. Vaccinate the health-care workers and the policemen in the state with the highest weighted positive testing rate given by $(\text{positive testing rate} \times R_0 \times \log(\text{Population}))$ where positive testing rate is the share of tests returning a positive result [25]
2. Step 1: Identify demographic clusters/nodes within the state on the basis of the age of the population
3. Step 2: Social interactions between the nodes and within them need to be modelled using edges with representative weights
4. Step 3: Once these nodes and edges are formed, an index would be created for each node which would essentially be

a weighted sum of all of its edges. $w_i = \begin{pmatrix} we_{i,1} \\ we_{i,2} \\ \vdots \\ we_{i,j} \end{pmatrix}$ where $we_{i,j}$ is the weight of the j th edge of the i th node

$$W_d = \sum_{i=1}^g \sum_{j=1}^n we_{i,j} = \sum_{i=1}^g w_i \tag{50.23}$$

where W is the weight vector representing the individual weights of g nodes.

- Step 4: Vaccinate each node in the proportion of its representative weight i.e $w_i / \sum w_i$. Thus the number of vaccines administered to each group would be represented as:

$$v_{i,d} = (w_i / \sum_{i=1}^g w_i) N(d), \tag{50.24}$$

where v_i represents the number of vaccines required for i th node and $N(d)$ is the total number of vaccines available for distribution on the d th day. Now these v_i vaccines are distributed to the i th node in the ratio of the income distribution in the same node. Assuming that the entire population of the node is divided into three economic groups : low income, middle income and high income such that:

$$p_{i,d} = \begin{pmatrix} pli_i \\ pmi_i \\ phi_i \end{pmatrix} \tag{50.25}$$

where pli_i, pmi_i, phi_i are the population of the low, middle and high income groups respectively in the i th node and the vaccines are distributed in the ratio $pli_i : pmi_i : phi_i$. P is the Population vector which is defined as:

$$P = (p_1, p_2, p_3 \dots p_g) \tag{50.26}$$

After completing v_i vaccinations, the population vector needs to be updated as follows:

$$p_{i,d} = p_{i,d-1} - (p_{i,d-1} / \sum p_{i,d-1}) v_{i,d-1} \tag{50.27}$$

- Step 5: Reduce the weights of all outgoing edges of the nodes that were vaccinated in the proportion of their respective v_i :

$$w_{i,d} = (p_{i,d} / p_{i,d-1}) \sum_{j=1}^n w_{i,j,d-1} \tag{50.28}$$

- Step 6: Recalculate indices for the nodes
- Step 7: Jump to Step 4

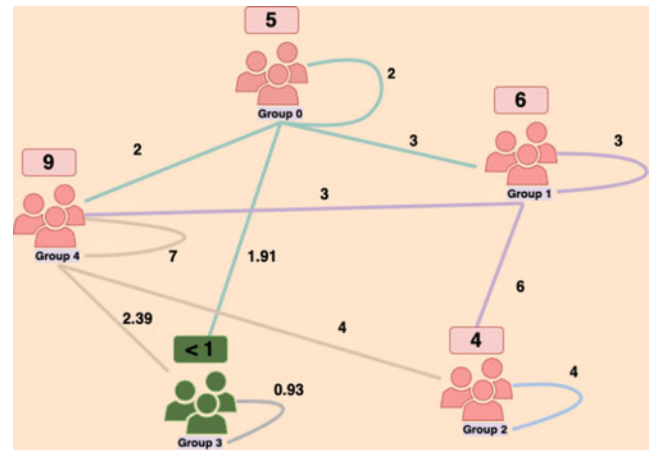


Fig. 50.7 Vaccine Distribution Algorithm

50.9 Tracking the Mutations of the Virus

Influenza viruses are known to change from year to year and hence vaccines designed to prevent them are expected to be updated annually to include the viruses most likely to circulate in the upcoming flu season. Once a decision has been made regarding the selection of viruses for the upcoming season, the manufacturers have to operate on a tight timeline for producing, testing, releasing and distributing the vaccine. Due to these strict time constraints, any problems encountered during production may cause shortages or delays in the supply of vaccines [26]. The COVID-19 vaccine development process should be expected to be subjected to similar constraints, although the virus seems to be mutating at a slower rate than normal influenza viruses; primarily due to the replication mechanism incorporated by corona-viruses which include a proof reading mechanism after replication that helps reduce the number of errors that accumulate with replication [27]. While mutations are impossible to predict, it's often suggested that the virus is likely to become less lethal with the progression of the pandemic. As the proportion of people who are immune to the disease increases, the virus is less likely to find a host to spread to. Therefore, in accordance to evolutionary theory, the less lethal strains of the virus which are less likely to kill the host, are likely to gain dominance. Although this thought has been disputed by those who have suggested evidence asserting the fact that how harmful the virus is to humans may not be as important a factor for determining the likelihood of its spread in the human population. Therefore it remains to be seen whether Sars-COV-2 becomes less lethal over time [28]. Mutations, by virtue of their very nature, are random events; so predicting when they will occur and the impact of a specific mutation on the harmfulness of the virus is an arduous task. However, monitoring the mutations of

the virus as it spreads is not just crucial for understanding the changes in its lethality or transmissibility, but also how they could affect the tools that are being developed in response.

50.10 Conclusion

In this paper, various mathematical and numerical analyses have been conducted to understand the spread of COVID-19 across India as a nation and her individual states. Our implementation of the time-dependent SIR model, with a mean error percentage of 1%, helped us analyse the impact of the social distancing policies imposed by the Indian government in a more robust and adaptive way than a classical SIR model would have. The results show that one-day prediction errors for the number of infected cases $x(t)$ and the number of recovered cases $z(t)$ are within $\approx 5\%$ for the dataset collected from <https://covid19india.org>. Our SIRV model was able to statistically demonstrate the impact of vaccination on the different states of the country; this could be used to assess and prioritise the vaccine distribution exercise. The novel vaccine distribution algorithm proposed in the paper has been developed with the intention of ensuring an equitable, efficient and effective vaccination campaign.

References

1. M.A. Ramsay, John Snow, MD: anaesthetist to the Queen of England and pioneer epidemiologist. *Proc (Bayl Univ Med Cent)* (2006)
2. M.F. McGuire, M.S. Iyengar, D.W. Mercer, Computational approaches for translational clinical research in disease progression. *J. Investig. Med.* **59**(6), 893–903 (2011)
3. C.L. Ventola, The antibiotic resistance crisis: Part 1: Causes and threats. *P. T* **40**(4), 277 (2015)
4. M.J. Binnicker, Challenges and controversies to testing for COVID-19. *J. Clin. Microbiol.* **58**(11) (2020)
5. H.J. Sussmann, On the gap between deterministic and stochastic ordinary differential equations. *Ann. Probab.*, 19–41 (1978)
6. T. Harko, F.S.N. Lobo, M.K. Mak, Exact analytical solutions of the susceptible-infected-recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Appl. Math. Comput.* **236**, 184–194 (2014)
7. F. Brauer, C. Castillo-Chavez, Basic ideas of mathematical epidemiology, in *Mathematical Models in Population Biology and Epidemiology*, (Springer, New York, 2001), pp. 275–337
8. J.D. Murray, *Mathematical biology: I. an introduction*. Vol. 17. (Springer Science & Business Media, 2007)
9. D.J. Daley, J. Gani, *Epidemic Modeling: An Introduction* (Cambridge University Press, Cambridge, 2005)
10. F. Brauer, P. van den Driessche, J. Wu (eds.), *Lecture Notes in Mathematical Epidemiology* (Springer, Berlin, Heidelberg, 2008), pp. 19–79
11. Y.-C. Chen, P.-E. Lu, C.-S. Chang, T.-H. Liu, A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering* **7**(4), 279–3294 (2020)
12. The Lancet, India under COVID-19 lockdown. *Lancet (London, England)* **395**(10233), 1315 (2020)
13. C.S. Pramesh, *COVID-19: The Lockdown Is Ending - But India's Epidemic Is Far From Over* (The Wire, 2020)
14. L. Thunström, S.C. Newbold, D. Finnoff, M. Ashworth, J.F. Shogren, The benefits and costs of using social distancing to flatten the curve for COVID-19. *Journal of Benefit-Cost Analysis* **11**(2), 179–195 (2020)
15. Guidelines on Clinical Management of COVID19, <https://www.mohfw.gov.in/>
16. S. Funk et al., Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western area region of Sierra Leone, 2014–15. *PLoS Comput. Biol.* **15**(2), e1006785 (2019)
17. World Health Organization. “Immunity Passports” in the context of COVID-19: scientific brief, 24 April 2020. WHO/2019-nCoV/Sci_Brief/Immunity_passport/2020.1. World Health Organization, 2020
18. S. Mukhopadhyay, D. Chakraborty, Estimation of undetected COVID-19 infections in India, in *medRxiv*, (2020)
19. M. Peiris, G.M. Leung, What can we expect from first-generation COVID-19 vaccines? *Lancet* **396**(10261), 1467–1469 (2020)
20. S. Gao, Z. Teng, J.J. Nieto, A. Torres, Analysis of an SIR epidemic model with pulse vaccination and distributed time delay. *J. Biomed. Biotechnol.* **2007** (2007)
21. R. Aguas, R.M. Corder, J.G. King, G. Goncalves, M.U. Ferreira, M. Gabriela, M. Gomes, Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics, in *medRxiv*, (2020)
22. The World Bank, <https://data.worldbank.org/indicator/SP.POP.TOTL>
23. M. Chaudhary, P.R. Sodani, S. Das, Effect of COVID-19 on economy in India: Some reflections for policy and programme. *J. Health Manag.* **22**(2), 169–180 (2020)
24. J. Corum, S.-L. Wee, C. Zimmer, Coronavirus vaccine tracker. *The New York Times* **5** (2020)
25. Coronavirus (COVID-19) Testing, <https://ourworldindata.org/coronavirus-testing> (2020)
26. Centers for Disease Control and Prevention, How the Flu Virus Can Change: “Drift” and “Shift” (2019)
27. F. Robson, K. Shahed Khan, T. Khanh Le, C. Paris, S. Demirbag, P. Barfuss, P. Rocchi, W.-L. Ng, Coronavirus RNA proofreading: Molecular basis and therapeutic targeting. *Molecular cell* (2020)
28. Virulence: A positive or negative trait for evolution? <https://www.virology.ws/2009/06/10/virulence-a-positive-or-negative-trait-for-evolution/>

Virtual Hospital: A System for Remote Monitoring of Patients with COVID-19

51

Vanessa Stangerlin Machado Paixão-Cortes, Walter Ritzel Paixão-Cortes, Dorval Thomaz, Felipe de Siqueira Zanella, Ricardo Luís Ravazzolo, and Gerson Luis da Silva Laureano

Abstract

Coronavirus disease represents a global public health concern. To minimize its damage is necessary to create technologies for the prevention and control of this emerging disease. However, there is still no equipment that can remotely monitor COVID-19 patients, regardless of where the patient is. Within the scope of the Internet of Things, this article presents the Virtual Hospital, a system for remote monitoring of patients with COVID-19. The device features biosensors that are configured to monitor the symptoms of COVID-19, taking into account the specification of the NEWS2 protocol, among them: temperature, heart rate, blood oxygenation, systolic pressure, level of consciousness, and respiratory rate. With the monitoring system, health professionals will continually monitor patients without the need for physical contact because any anomaly that may occur will cause the system to notify the person responsible for taking appropriate action immediately. We believe that, as a consequence, its use can decrease the number of critically ill patients in hospitals, reducing in-hospital mortality rates.

Keywords

COVID-19 · Technologies · Prevention · Monitoring · Internet of things · NEWS2 protocol · Pandemic · SOA · LoraWan · SigFox

V. S. M. Paixão-Cortes (✉) · W. R. Paixão-Cortes · D. Thomaz
F. de Siqueira Zanella · R. L. Ravazzolo · G. L. da Silva Laureano
Universidade Federal de Ciências da Saúde de Porto Alegre, DECESA,
Porto Alegre, Brazil
e-mail: vanessapc@ufcspa.edu.br

51.1 Introduction

Coronaviruses are a large family of viruses, where some are responsible for causing disease. These diseases can be less severe, like the common cold, or more serious, like MERS (Middle East Respiratory Syndrome) and SARS (Severe Acute Respiratory Syndrome), and can be easily transmitted from person to person, while other diseases cannot. In December 2019, Chinese authorities determined a new coronavirus identified in a hospitalized person [1]. Afterward, a local outbreak of pneumonia occurred, caused by this new pathogen, called coronavirus 2 from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2].

Since then, the outbreak has spread to all provinces in mainland China and 27 other countries and regions [2]. Coronavirus disease, called COVID-19, is caused by SARS-CoV-2 and represents the causative agent of a potentially fatal disease that is a major global public health problem [3]. The COVID-19 infection has spread, and its incidence has increased worldwide [4]. The first deaths occurred mainly in the elderly, among whom the disease can progress more quickly [4]. According to the World Health Organization, in November 2020, there were more than 53.7 million confirmed cases and 1.3 million deaths from the pandemic [5].

Person-to-person transmission of COVID-19 infection led to the isolation of patients who were subsequently subjected to a variety of treatments, as well as extensive measures to reduce the contagion of COVID-19 from person to person [3]. Given the above scenario, there is a need for special attention to monitoring, protecting, or reducing the transmission of the disease, where new technologies must be created for susceptible populations, including children, health professionals, and the elderly [3]. Helping public health professionals to recognize and deal with the new coronavirus quickly, effectively, and calmly, with an updated understanding, is essential to containing the pandemic [4].

In response to this public health emergency, this article aims to present a prototype of a remote monitoring system for COVID-19 symptoms, called Virtual Hospital, using IoT (Internet of Things) and biosensors. Following the NEWS2 protocol (National Early Warning Score 2), the signs of COVID-19 are monitored—current temperature, heartbeat, blood oxygenation, cough, and shortness of breath. The hypothesis is that an IoT remote monitoring system may favor patients' follow-up with COVID-19, including in patients' homes, to prevent them from seeking help only in the most severe cases of the disease.

51.2 Remote Patient Monitoring Systems

Notably, in the last 10 years, medicine has evolved mainly concerning technological issues, and currently, it is the primary driver of the technological innovation market [6]. Health care is transforming. It is imperative to leverage new technologies to generate new data and support the advent of precision medicine, as recent scientific and technological advances have improved our understanding of disease pathogenesis and changed how we diagnose and carry out the treatment. This contributes to obtaining more accurate, predictable, powerful, and personalized healthcare for each patient [7].

Within this perspective, in the clinical treatment of diseases, especially in COVID-19, there is a need to monitor patients even more, due to the number of older people, with chronic illnesses or critical health conditions. This scenario causes a high demand, demanding more and more of the current infrastructure of health services. In this sense, remote patient monitoring systems have been widely disseminated to different medicine areas [8–10].

However, pervasive home health care applications require a specialized hardware and software infrastructure capable of collecting and processing patient and environment data [8]. In this sense, remote monitoring has undergone significant developments, as the advent of IoT supports that low-cost solutions can be safely implemented [9–11]. IoT occurs when the internet and networks are combined, and depending on the infrastructure, it can be thought of as the combination of embedded electronics, sensors, software, and connectivity. The data can be accessed using resources provided by the internet from anywhere and at any time [9].

The cloud and IoT, and mobile technologies facilitate the monitoring of patient's health conditions, sharing health information with health teams [9]. Sensors and technological equipment that measure patient parameters make us believe it is possible to reduce losses concerning human resources and infected patients. In fact, through these technologies, it is possible to provide the health professional with a complete overview of the patient's clinical condition, even without hav-

ing visual contact, through data collected by biosensors and analyzed in the cloud by artificial intelligence algorithms, as well as “deep” and “machine” learning. Biosensor data is sent to the “cloud” and allows it to be made available and read in real-time. The information is stored and transmitted to a supervisory center capable of handling emergencies. This control can favor the reduction of deaths and contamination rates.

Besides, one of the advantages of adopting monitoring systems is the ease of exchanging information between the patient and the doctor without the need for physical contact between them, which allows the release of hospital beds, as patients can remain in their households with fast service in case of emergency [12]. Patients present “clues,” which are perceptible, hours before the condition becomes exceptionally severe [13]. In this sense, monitoring the patient with COVID-19 is crucial since some patients deteriorate their health condition in 2 weeks, most commonly due to pneumonia [14].

Our proposal presents the differential that Virtual Hospital, in addition to easily monitoring the symptoms of COVID-19, can be used both in a hospital and in a residential environment utilizing national technology, that is, Brazilian technology, which was used in its implementation.

51.3 Symptoms Monitoring and COVID-19

According to Greenhalgh et al. [14] most patients with COVID-19 can be managed remotely, with advice on symptom management and self-isolation. The success of care for acutely ill patients will, of course, depend on their correct diagnosis. However, with the diagnosis identified early, the measures implemented will be more effective, which will lead to a better prognosis with a lower financial cost. Thus, simple measures of low economic impact and great effectiveness could change patients' evolution if they were taken at the right time. Failure to recognize complications early leads to worsening of the disease; in many cases, it makes the response slow and does not always succeed. For this reason, health institutions must improve the ability to identify critically ill patients early [13].

For this, there are alert scales. The NEWS (National Early Warning Score) is an alert scale based on a weighted system of scoring vital parameters. Its primary purpose is the early identification of the risk of acute deterioration of the patient [15]. Its main objective is to ensure early care by identifying signs of the clinical worsening of the patient, standardizing urgency and emergency care in patients admitted to the open unit, even in the absence of the attending physician, to reduce the incidence of Cardiac Arrest Breaks and decrease in-hospital mortality. This favors the safety of the patient, the multidisciplinary team, and the Institution [15].

As reported by EBSERH (Empresa Brasileira de Serviços Hospitalares) [13] NEWS uses physiological parameters to obtain a score, which increases according to the change in the normal range. We can check the parameters in Fig. 51.1.

In their applicability context, physiological parameters are evaluated, each parameter receiving a score of 0 to 3 points. The score is defined based on the sum of the scores achieved in the evaluation, such as temperature, heart rate, systolic blood pressure, peripheral oxygen saturation, oxygen supplementation [13].

The NEWS model was used by the Government of Paraíba, PB, Brazil, through the SES-PB network, in line with the Central Regulation Center (CEHR) of COVID-19, which developed a score, more simplified and short, to direct the clinical decision and the type of management to be indicated in patients, since the majority of COVID-19 infections are self-limiting. The most critical cases are commonly associated with the elderly population and individuals with comorbidities.

In China, Liao et al. [17] developed the modified NEWS score for patients with COVID-19, with the addition of age, as a factor that changes the score. All other items on the score were maintained, with the same original formula [18]. The Government of Paraíba is already using the modified NEWS score to monitor patients with COVID-19, called NEWS-FAST-COVID, considering the most common symptoms like fever (>37.8), dry cough, dyspnea, myalgia, runny nose, fatigue in the last 7 (seven days), and the reality of Brazilian hospitals [18].

Depending on the score obtained, guidelines are sent to patients. Without alarm signs, after evaluation, the patient is referred to home isolation, advised to seek health services, to be observed for 6 h to 24 h, in a hospital UPA, with Laboratory Image diagnosis if possible or referred for immediate medical conduct, evaluating the ICU vacancy, or COVID-19 Reference Center [18].

Fig. 51.1 National Early Warning Score (NEWS), showing the list of parameters and how they will be scored according to values. The score is between 0 and 3, with 0 being neutral and 3 should trigger an alert [16]

PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U

51.4 Methodology

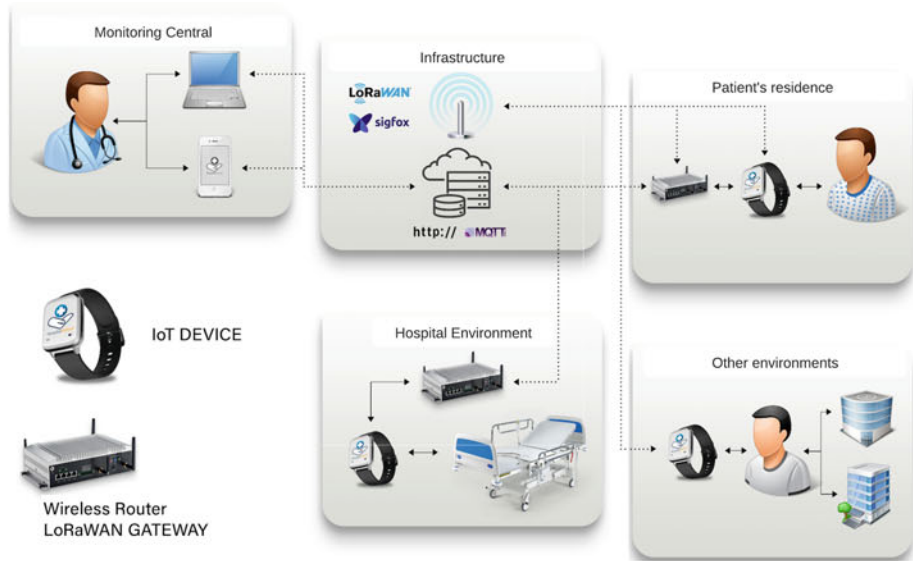
This section identifies how the Virtual Hospital was developed, and the resources needed to achieve the proposed objective. Our main objective is to enable the development of a remote monitoring system that allows the identification and parameterization of symptoms according to the NEWS2-COVID-19 protocol, warning about the patient’s acute clinical deterioration.

Agile Software Development and Design Thinking techniques were used to develop the Virtual Hospital system. This process was interactive and included the following steps:

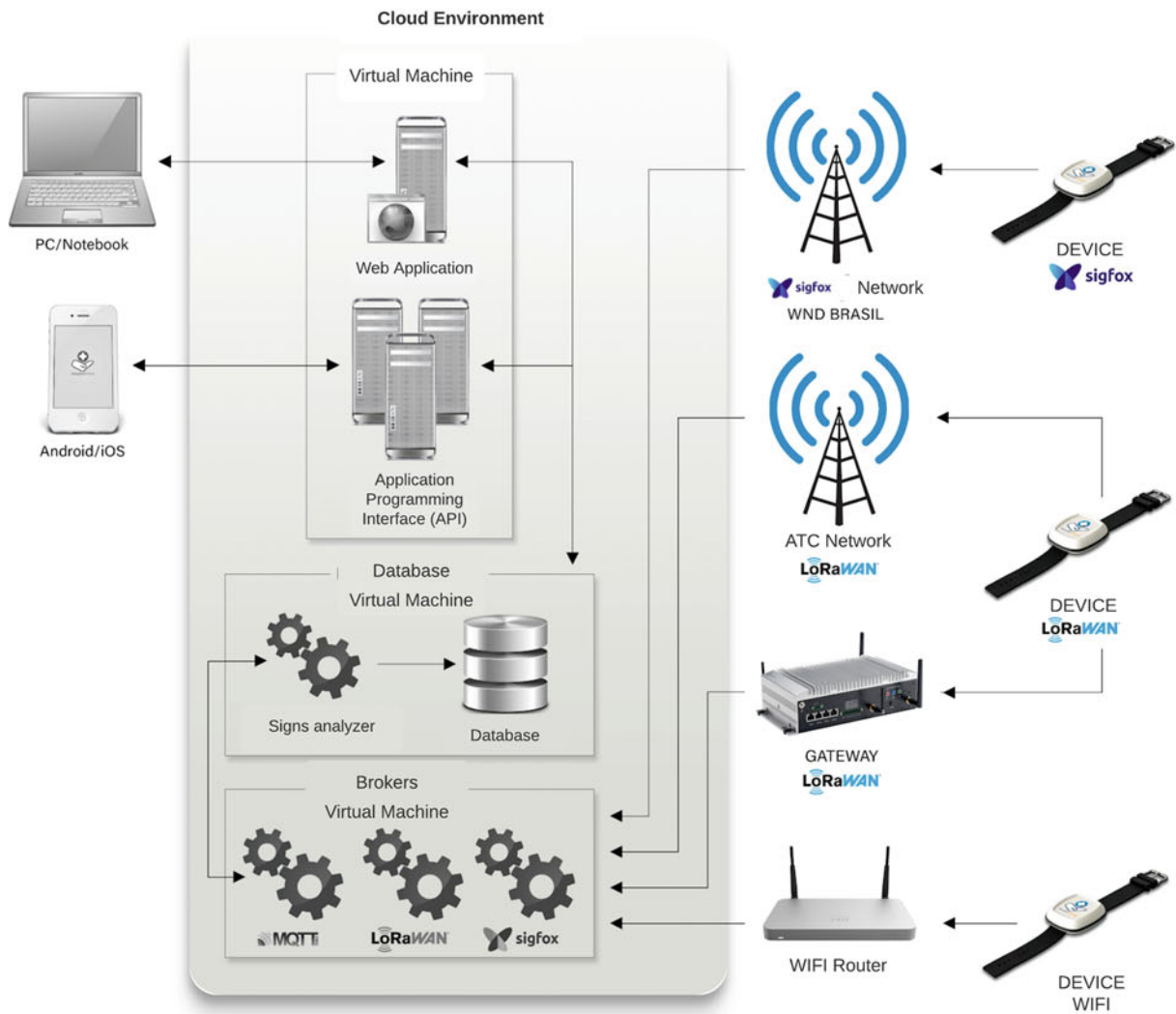
- Context analysis: understand the context of use of Remote Patient Monitoring (MRP), IoT technologies, and biosensors. It involved literature research on IoT and biosensor technologies, as well as programming languages and other related technologies.
- Needs analysis and planning: mapping of technologies that could be used to develop the MRP system.
- Specification and design of interface: specification of requirements, design of the environment, and validation by building proofs of concept tested with simulated patients.

51.5 Results

In this section, we present the functionalities and technologies used in developing the Virtual Hospital MRP system 51.2. This study developed a monitoring center that identifies the platform where healthcare professionals or authorized persons are registered. Access can be performed through web and mobile interfaces, as long as the devices that allow the connection. The infrastructure environment Fig. 51.2a, which presents the hosting of the system in cloud



(a)



(b)

Fig. 51.2 Clinical MRP—remote monitoring of patients. (a) Infrastructure. (b) Service-oriented architecture

computing, where the data collected from the IoT devices are received, processed, stored, and distributed.

For the other environments shown in Fig. 51.2b, the patient can be in three types of locations, hospital, residential, and others. It is understood as residential, where the patient is fulfilling social isolation outside an ICU (Intensive Care Unit) or hospital bed, sending data to the cloud through a wireless router or LoRa gateway, which serve as a means of communication between the system cloud computing and the IoT device. The hospital environment is understood to be a public or private health unit where the patient can be served either in infirmary or ICU beds, using the same technologies used in the residential environment as a means of communication.

In monitoring situations involving low power consumption and long-range wireless mobile devices, embedded IoT technologies are used. Considering the potential that these technologies bring to our daily lives, the product presented in this work is an IoT solution that uses LoRaWAN, Sigfox, wi-fi, and Bluetooth 5.0 technologies.

In this study, the prototype used biosensors, data storage, and a web application using programming languages, such as Python, Flutter, and Angular, to remotely view the patient's condition. To use this wearable biosensor, it is essential to define the parameters that will be measured with the patient. The MRP infrastructure is based on a Service-Based Architecture (SOA) composed of virtual machines, data networks, and remote devices. The communication between these components uses encryption protocols, thus ensuring a high security level over the data that travels over the network and stored in the database.

Figure 51.2 presents a complete MRP (Virtual Hospital) map. In this Cloud Services model, the use of three types of virtual machines is proposed: for Brokers who receive measurements from patients; for the Patient Measurement Database and Analyzer Service; and for the Application Programming Interface and Web Application.

- **Virtual Machine Brokers:** brokers are responsible for receiving the data collected on the sensors and making them available to the Patient Measurement Analyzer Service, without performing any processing.
- **Database Virtual Machine:** runs the Database services and the Patient Measurement Analyzer Service.
- **Patient Measurement Analyzer Service:** responsible for reading the patient measurements received from the Brokers and for processing this data. The measurements are analyzed and compared with the NEWS2 protocol table, thus generating a patient's score. Depending on the result, the analyzer service triggers an alert in the system, indicating the patient's current state.
- **Database:** The Database Management System (DBMS) is responsible for storing all MRP data available to the

Application Programming Interface (API). The DBMS uses Transparent Data Encryption technology (TDE) to store data. The Database was implemented in a cluster structure, formed by a set of virtual machines: a primary one that will replicate all data to a secondary machine, providing a high availability environment.

- **Application Virtual Machine:** where API and web application services are executed. The API is responsible for mediating the access of web and mobile applications to the Database. In the API, we have the implementation of information access security, as well as the other business rules related to the data. The advantage of using this type of service is that it allows multiple interfaces to access the same logic, thus ensuring the integrity of the information being manipulated.
- **Web Application—**The web application is responsible for the procedures for maintaining patient records, sensors, types of alerts, NEWS2 score table, among others. The web application also provides all monitoring reports and a monitoring dashboard for active patients.
- **Mobile Application—**has the same functionality as the web application, allowing maintenance and monitoring procedures to be carried out. It is installed and used only on mobile devices with Android or iOS systems.
- **Secure Communication among services:** Communication between all services and components of the solution is encrypted, using the SSL/TLS (Secure Socket Layer), LoRaWAN, Sigfox, and Digital Certificates protocols.

It should be noted that this is a scalable environment so that new virtual machines of the same type can be added, forming a cluster or a set of machines, providing greater processing power and exceptionally high availability of services. The biggest difficulty faced during the implementation of the solution was related to the tests with the devices in LoRaWAN and Sigfox networks, since the team does not control these.

51.6 Conclusion

The development of the Virtual Hospital solution¹ was motivated by the need to obtain positive results concerning the reduction of the virus's spread well as a better optimization of hospital beds, making use of hospitalization or transfer of unit only when necessary.

The access to remote monitoring technology, with data collection in a fully automatic way, opens the possibility of providing care to infected patients more assertively and simultaneously. Our main contributions are described below:

¹<http://www.hospitalvirtual.com.br/>.

- COVID-19 remote monitoring system: based on the NEWS2 protocol, the system will be able to monitor patient's symptoms easily at home or hospital environments.
- Adherence to the NEWS2 protocol: based on this stored data, the health monitoring unit can obtain live information and decide about a patient.
- Creation of technological innovation with immediate application in the treatment and control of the spread of Coronavirus, with intellectual and technological properties that will serve as a basis for future studies in the area of Medical technology.

The Virtual Hospital project, even in times of pandemic with 100% of remote work, has already won some awards in 2020: Hack for Brazil 2020 (Top 4) and Ideathon COVID-19 (first place). As future work, the system will be tested with end-users through a partnership already consolidated with a health plan company in Brazil.

References

1. WHO, WHO statement regarding cluster of pneumonia cases in Wuhan, China, Jan 2020 [Online]. Available: <https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china>
2. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**(5), 533–534 (2020)
3. H.A. Rothan, S.N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J. Autoimmun.* **109**, 102433 (2020)
4. W. Wang, J. Tang, F. Wei, Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. *J. Med. Virol.* **92**(4), 441–447 (2020)
5. WHO, Coronavirus disease (COVID-19) situation report. Data as received by who from national authorities, as of 10am cest 15 November 2020. Nov 2020 [Online]. Available: <https://www.who.int/publications/m/item/weekly-epidemiological-update---17-november-2020>
6. E.J. Topol, A decade of digital medicine innovation. *Sci. Transl. Med.* **11**(498), eaaw7610 (2019)
7. A.A. Seyhan, C. Carini, Are innovation and new technologies in precision medicine paving a new era in patients centric care? *J. Transl. Med.* **17**(1), 114 (2019)
8. S.T. Carvalho, M. Erthal, D. Mareli, A. Sztajnberg, A. Copetti, O. Loques, R. de Janeiro-RJ-Brasil, Monitoramento remoto de pacientes em ambiente domiciliar, in *XXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos-Salao de Ferramentas, Gramado, RS* (2010), pp. 1005–1012
9. A.M. Ghosh, D. Halder, S.A. Hossain, Remote health monitoring system through IoT, in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)* (IEEE, New York, 2016), pp. 921–926
10. M. Hassanaliheragh, A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, B. Kantarci, S. Andreeescu, Health monitoring and management using internet-of-things (IoT) sensing with cloud-based processing: opportunities and challenges, in *2015 IEEE International Conference on Services Computing* (IEEE, New York, 2015), pp. 285–292
11. W.P. de Silveira Junior, L.G.L. Moura, Comunicação IoT aplicado à saúde através de dispositivos de monitoramento pessoal. *LINKSCIENCEPLACE-Interdiscipl. Sci. J.* **5**(3), 1–5 (2019)
12. A. Machado, E.L. Padoin, F. Salvadori, L. Righi, D.M. Campos, P.S. Sausen, S.L. Dill, Utilização de dispositivos móveis, web services e software livre no monitoramento remoto de pacientes, in *Congresso Brasileiro de informática na saúde, XI. Anais* (2008)
13. H.U.F. EBSERH, Escore para alerta precoce-ebserh [governamental], ministério da educação, brasil. Nov 2020 [Online]. Available: <http://www2.ebserh.gov.br/web/hu-ufjf/escore-para-alerta-precoce>
14. T. Greenhalgh, G.C.H. Koh, J. Car, COVID-19: avaliação remota em atenção primária à saúde. *Revista Brasileira De Medicina De Família E Comunidade* **15**(42), 2461–2461 (2020)
15. R. Gomes, Implementando protocolo de deterioração aguda precoce (news e meows) e o papel da equipe multidisciplinar [palestra] iv seminário de segurança do paciente no ambiente hospitalar, hospital jorge valente, 2026 [Online]. Available: <http://www.cremeb.org.br/wp-content/uploads/2016/07/Mesa-02-Palestra-03-Rodrigo-Silva-Gomes.pdf>
16. B. William, G. Albert, C. Ball, D. Bell, R. Binks, L. Durham, J. Eddleston, N. Edwards, D. Evans, M. Jones et al., National early warning score (news): standardizing the assessment of acute illness severity in the NHS. Report of a Working Party (2012)
17. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K.S. Leung, E.H. Lau, J.Y. Wong et al., Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New Engl. J. Med.* **382**, 1199–1207 (2020)
18. G.D.E. Paraiba, Estratificação de risco de criterios de internamento em uti, April 2020 [Online]. Available: <https://paraiba.pb.gov.br/diretas/saude/coronavirus/evidencias-cientificas/arquivos/protocolo-news-fast-covid-19.pdf>

Single-Cell RNA Sequencing Data Imputation Using Deep Neural Network

52

Duc Tran, Frederick C. Harris Jr., Bang Tran, Nam Sy Vo, Hung Nguyen, and Tin Nguyen

Abstract

Recent research in biology has shifted the focus toward single-cell data analysis. The new single-cell technologies have allowed us to monitor and characterize cells in early embryonic stage and in heterogeneous tumor tissue. However, current single-cell RNA sequencing (scRNA-seq) technologies still need to overcome significant challenges to ensure accurate measurement of gene expression. One critical challenge is to address the dropout event. Due to the low amount of starting material, a large portion of expression values in scRNA-seq data is missing and reported as zeros. These missing values can greatly affect the accuracy of downstream analysis. Here we introduce scIRN, a neural network-based approach, that can reliably recover the missing values in single-cell data and thus can effectively improve the performance of downstream analyses. To impute the dropouts in single-cell data, we build a neural network that consists of two sub-networks: imputation sub-network and quality assessment sub-network. We compare scIRN with state-of-the-art imputation methods using 10 scRNA-seq datasets. In our extensive analysis, scIRN outperforms existing imputation methods in improving the identification of cell sub-populations and the quality of visualizing transcriptome landscape.

D. Tran · F. C. Harris Jr. (✉) · B. Tran · H. Nguyen · T. Nguyen
Computer Science & Engineering, University of Nevada, Reno, Reno, NV, USA

e-mail: duct@nevada.unr.edu; fred.harris@cse.unr.edu;
bang.t.s@nevada.unr.edu; hungnp@nevada.unr.edu; tinn@unr.edu

N. S. Vo

Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam

e-mail: v.namvs@vintech.net.vn

Keywords

Single cell · scRNA-seq · Imputation · Sequencing · Neural network · Gene expression · Residual network · Dimension reduction · Clustering · Visualization

52.1 Introduction

The ability to monitor and characterize biological samples at single-cell resolution has opened up many novel research fields, such as studying cells in early embryonic stage or decomposition heterogeneous environment of cancer tumor [1, 2]. These promising applications have led to the generation of a massive amount of single-cell data, where each dataset consists of hundreds of thousands of cells [3–5].

Current single-cell RNA sequencing (scRNA-seq) technologies still need to overcome significant challenges to ensure the accurate measurement of gene expression [6, 7]. One notable challenge of scRNA-seq is the dropout events, which happen when a highly expressed gene has no expression value in the sequencing data [8]. The sources of these errors can be attributed to the limitation of sequencing technologies. Due to the low amount of starting mRNA collected from individual cells, failed amplification can happen and causes the expression values to be inaccurately reported [9–11]. This leads to an excessive amount of zeros in the expression values of scRNA-seq data. On the other hand, the zero expression values can also be due to biological variability. Since downstream analyses of scRNA-seq are performed on gene expression data, it is essential to have a precise expression measurement. Therefore, imputing scRNA-seq data to recover the information loss caused by dropout would greatly improve the quality of downstream analyses.

To address the dropout challenge, a number of imputation methods have been developed to infer the missing data [12–19]. Those methods can be classified into two

categories: (1) statistical-based methods, and (2) diffusion smooth-based methods. Methods in the first category include bayNorm [12], SAVER [13], scImpute [14], scRecover [15], and RIA [17]. These methods typically model the data as a mixture of distributions. For example, scImpute models the gene expression as a mixture of two different distributions: the Gaussian distribution represents the actual gene expression while the Gamma distribution accounts for the dropout events. Similarly, SAVER [13] models read counts as a mixture of Poisson-Gamma and then uses a Bayesian approach to estimate true expression values of genes by borrowing information across genes. Another method, scRecover [15], uses the zero-inflated negative binomial model (ZINB) [20] to identify genes with zero-inflated expression. After identifying genes with true dropout, it uses the existing imputation methods such as scImpute, SAVER or MAGIC to impute the data. A more recent method, RIA [17], assumes that highly expressed genes follow a normal distribution and apply hypothesis testing method to identify true dropouts. Next, it imputes their values by using linear regression model. All of these methods assume the gene expression data follows a specific distribution, which does not always hold true in the reality. In addition, existing methods involve the estimation of many parameters for all genes across the whole genome. This can potentially lead to overfitting and high time complexity.

Methods in the second category include DrImpute [16], MAGIC [18], and kNN-smoothing [19]. MAGIC imputes zero expression values using a heat diffusion algorithm [21]. It constructs the affinity matrix between cells using Gaussian kernel and then constructs a Markov transition matrix by normalizing the sc-RNA similarity matrix. Next, MAGIC estimates the weights of other cells using the transition matrix. Another method is DrImpute [16] that is based on the cluster ensemble [22] and consensus clustering [23, 24]. It performs clustering for a predefined number of times and imputes the data by averaging expression values of similar cells. If the number of clusters is not provided by users, DrImpute uses some default values that might not be optimal for the data. kNN-smoothing is designed to reduce noise by aggregating information from similar cells (neighbors). The method assumes that the zero counts of scRNA-seq data follows a Poisson distribution. For cells that contain zero counts, kNN-smoothing performs a smoothing step using each cell's k nearest neighbors either through the application of diffusion models or weighted sums respectively. The major drawback of these methods is that they rely on many parameters to fine-tune their model, which often leads to over-smoothing the data.

Here we propose a new approach, single-cell Imputation using Residual Network (scIRN), that can reliably impute missing values from single-cell data. Our method consists of two steps. The first step is to generate a compressed and

accurate low-dimensional representation of the original data. The second step is to estimate the missing values using a neural network and information from the low-dimensional representation. The approach is tested using 10 single-cell datasets in comparison with four other methods. We demonstrate that scIRN outperforms existing imputation methods (MAGIC [18], scImpute [14], SAVER [13], and DrImpute [16]) in improving the identification of cell sub-populations and the quality of biological landscape.

52.2 Methods

The input of scIRN is an expression matrix, in which rows represent cells and columns represent genes or transcripts. The overall workflow of scIRN is described in Fig. 52.1, which consists of two modules: (1) generating a low-dimensional, non-redundant representation of the original data, and (2) imputing the dropout values. The purpose of the first module is to remove redundant signals and noise from the data. The output of the first module is a low-dimensional, non-redundant representation of the original data. This presentation is used as the target for the second module. In the second module, we impute the original data using a residual network. The parameters of the residual network are repeatedly adjusted so that the compressed representation of the imputed data is as similar to the non-redundant representation as possible. The details of each step are described in the following sections.

52.2.1 Generating Low-Dimensional, Non-redundant Representation

To generate a compressed, low-dimensional representation of original data, we apply our previously developed method, called scDHA [25]. scDHA consists of two core modules. The first module is a non-negative kernel autoencoder that can filter out genes or components that have insignificant contributions to data representation. The second module is a Stacked Bayesian Self-learning Network that is built upon the Variational Autoencoder [26] to project the filtered data onto a much lower-dimensional space. The output of scDHA is a low-dimensional matrix that preserves the global structure of the original data. This representation is used as the training target for the imputation module.

52.2.2 Imputing Dropout Data Using Residual Network

To impute the dropouts in single-cell data, we build a neural network that consists of two sub-networks. The first network

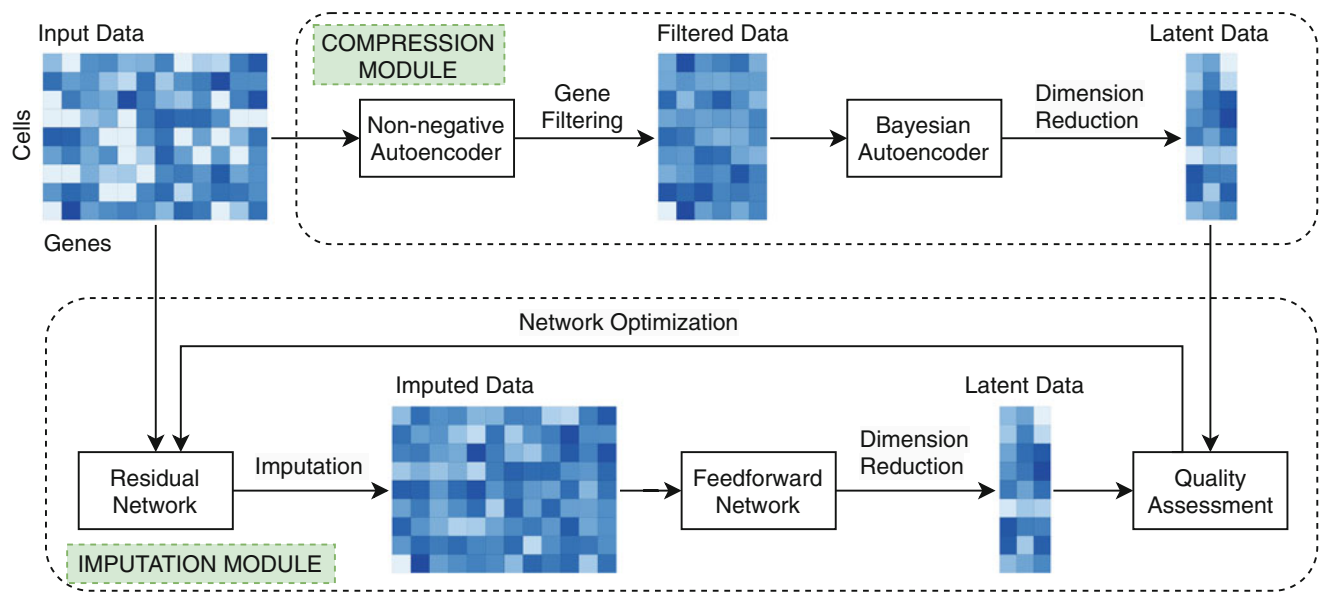


Fig. 52.1 The overall workflow of single-cell Imputation using Residual Network (scIRN). The first module (compression module) generates a compressed, low-dimensional representation of original data. The input data is first filtered (using an one-layer, non-negative kernel autoencoder) to remove genes that have insignificant contribution to the global structure of the data. After that, we project the data into a low-dimensional space to obtain a compressed data matrix (latent data).

This latent data is used as the training target for the imputation process. In the second module (imputation module), zero values in input matrix are imputed using a neural network-based imputation model. These imputed values are added to original data without modifying the non-zeros values to produce the imputed data matrix. The imputed data is compressed to a low-dimensional space (latent data). The parameters of the imputation module is repeatedly optimized by minimizing the difference between the two latent matrices

aims to infer the true value of zeros in the data. The output is a matrix with the same size as the input, in which the values at zero positions are modified. The non-zero values remain the same as of the original data. The second network aims to compress the imputed data to a lower dimension. This compressed data has the same size as the representation generated in the first step. By minimizing the difference between the representation generated from imputed data and the representation from the first step, the imputed values are ensured to have high accuracy.

The formulation of the neural network can be written as:

$$\begin{aligned} X_I &= f_I(X) \\ Z' &= f_C(X_I) \end{aligned}$$

where $X \in R_+^n$ is the input of the model (X is simply the original data), f_I and f_C represent the transformation by the two sub-networks, f_I imputes the zero values in the data, f_C compresses the imputed data onto a lower-dimensional space, and $Z' \in R^m$ ($m \ll n$) is the compressed data. For the f_I transformation, we use residual network [27] for a more stable and accurate imputation process. The network is optimized by minimizing $\|Z' - Z\|_2^2$, where Z is the low-dimensional representation generated by scDHA.

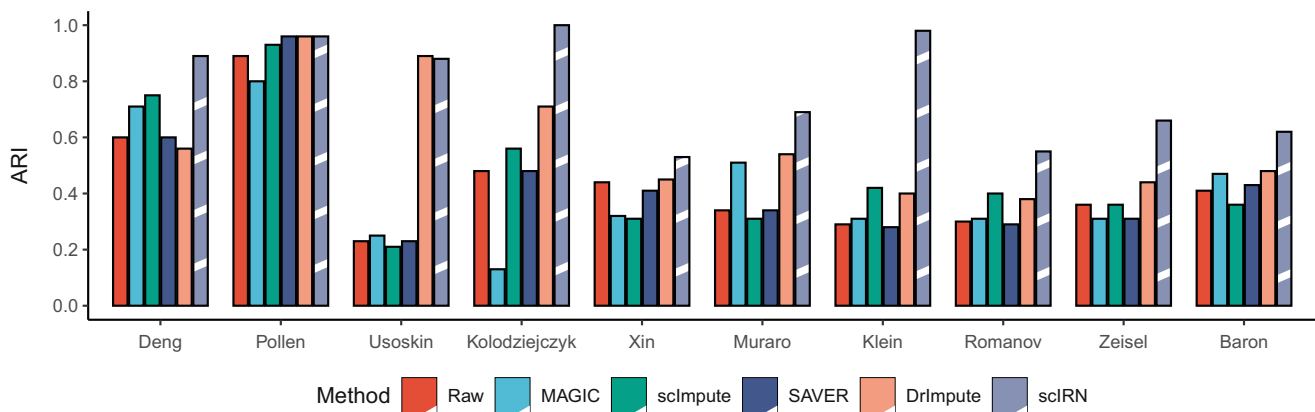
52.3 Results

We compares our method with four state-of-the-art imputation methods: MAGIC [18], scImpute [14], SAVER [13], and DrImpute [16]. Each of these methods represents a distinct strategy to single-cell data imputation: MAGIC is a Markov-based technique, DrImpute integrates clustering result from other software, while scImpute and SAVER use statistical models. Table 52.1 shows the 10 datasets used in our data analysis. The processed datasets were downloaded from Hemberg lab's website (<https://hemberg-lab.github.io/scRNA.seq.datasets>). In each dataset, the cell sub-populations are known. We used this information *a posteriori* to assess how the imputation methods improve the identification of cell populations, and how they enhance the visualization of transcriptome landscapes.

For each dataset, we used the above methods to impute the data. The quality of the imputed data is assessed using two downstream analyses, clustering and visualization. For clustering, we partitioned the data using k-means and compared the obtained partitioning against the true cell types using Adjusted Rand index (ARI) [37]. For visualization,

Table 52.1 Description of the 10 single-cell datasets used to assess the performance of imputation methods

Dataset	Tissue	Size	Class	Protocol	Accession ID	Reference
1. Deng	Mouse embryo	268	6	Smart-Seq2	GSE45719	Deng et al. [2]
2. Pollen	Human tissues	301	11	SMARTer	SRP041736	Pollen et al. [28]
3. Usoskin	Mouse brain	622	4	STRT-Seq	GSE59739	Usoskin et al. [29]
4. Kolodziejczyk	Mouse embryo stem cells	704	3	SMARTer	E-MTAB-2600	Kolodziejczyk et al. [30]
5. Xin	Human pancreas	1600	8	SMARTer	GSE81608	Xin et al. [31]
6. Muraro	Human pancreas	2126	10	CEL-Seq2	GSE85241	Muraro et al. [32]
7. Klein	Mouse embryo stem cells	2717	4	inDrop	GSE65525	Klein et al. [33]
8. Romanov	Mouse brain	2881	7	SMARTer	GSE74672	Romanov et al. [34]
9. Zeisel	Mouse brain	3005	9	STRT-Seq	GSE60361	Zeisel et al. 2015 [35]
10. Baron	Human pancreas	8569	14	inDrop	GSE84133	Baron et al. [36]

**Fig. 52.2** Adjusted Rand index (ARI) obtained from clustering on raw data and data imputed by MAGIC, SAVER, scImpute, DrImpute, and scIRN. The *x*-axis shows the names of the datasets while the *y*-axis

shows ARI value of each method. scIRN outperforms other methods in all datasets except Usoskin

we used UMAP [38] to generate the 2D representation and then calculated the silhouette index (SI) [39] of the 2D representation. SI measures the cohesion among cells of the same type, as well as the separation between different cell types.

52.3.1 scIRN Improves the Identification of Sub-populations

Given a dataset, we used the five methods to impute the data. After imputation, we have 6 matrices: the raw data and five imputed matrices (from MAGIC, scImpute, SAVER, DrImpute, and scIRN). To assess how separable the cell types in each matrix is, we reduced the number of dimensions using PCA and then clustered the data using k-means. The accuracy of cluster assignments is measured by ARI.

Figure 52.2 shows the ARI values for the raw and imputed data. Existing methods improve cluster analysis in some datasets but decreases the ARI values in some others. For example, MAGIC has higher ARIs than the raw data for the Deng, Usoskin, Muraro, Klein, Romanov, and Baron but has lower ARIs in the remaining 4 datasets. scIRN is

the only method able to improve the clustering performance compared to raw data in every dataset. Moreover, scIRN has the highest ARIs in all but Usoskin datasets. The average ARI of scIRN-imputed data is 0.77, which is higher than those obtained from raw data and data imputed by MAGIC, scImpute, SAVER, DrImpute (0.44, 0.41, 0.46, 0.43, 0.58, respectively).

For a more comprehensive analysis, we also report the assessment using normalized mutual information (NMI) and Jaccard index (JI) in Figs. 52.3 and 52.4, respectively. Regardless of the assessment metrics, scIRN outperforms other methods by having the highest NMI (10/10 datasets) and JI (9/10 datasets) values. These results demonstrate that cluster analysis using scIRN-imputed data leads to a better accuracy than using the raw data or data imputed by other imputation methods.

52.3.2 scIRN Improves Transcriptome Landscape Visualization

In this section, we demonstrate that scIRN improves the visualization of the single-cell data. We used UMAP [38] to

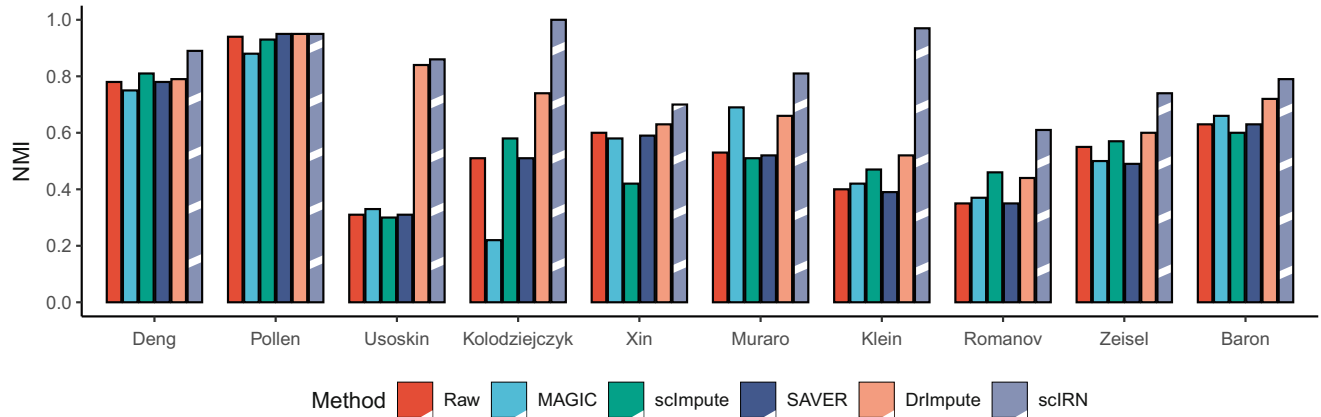


Fig. 52.3 Normalized mutual information (NMI) obtained from clustering on raw data and data imputed by MAGIC, SAVER, scImpute, DrImpute, and scIRN. The *x*-axis shows the names of the datasets while the *y*-axis shows NMI value of each method. scIRN outperforms other methods in all datasets

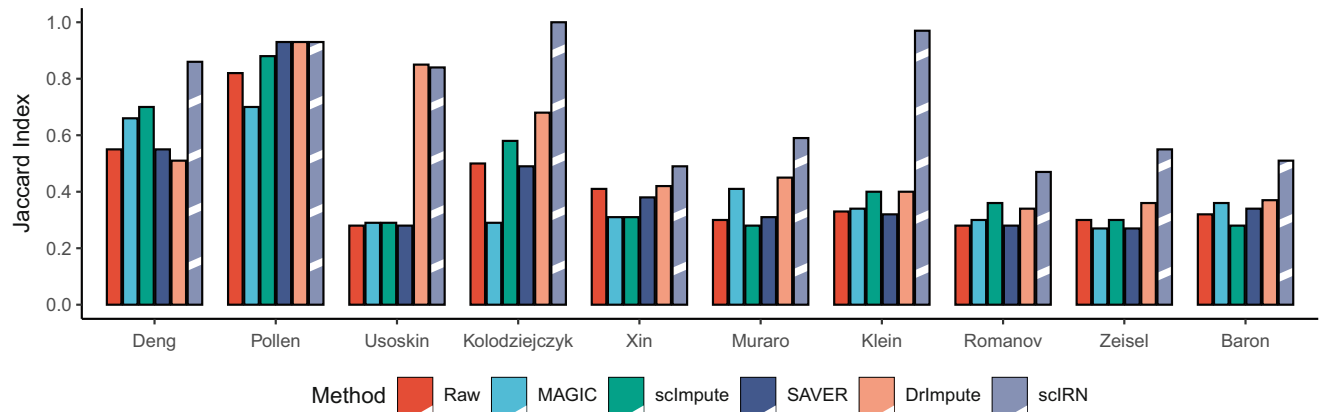


Fig. 52.4 Jaccard index (JI) obtained from clustering on raw data and data imputed by MAGIC, SAVER, scImpute, DrImpute, and scIRN. The *x*-axis shows the names of the datasets while the *y*-axis shows JI value of each method. scIRN outperforms other methods in all datasets except Usoskin

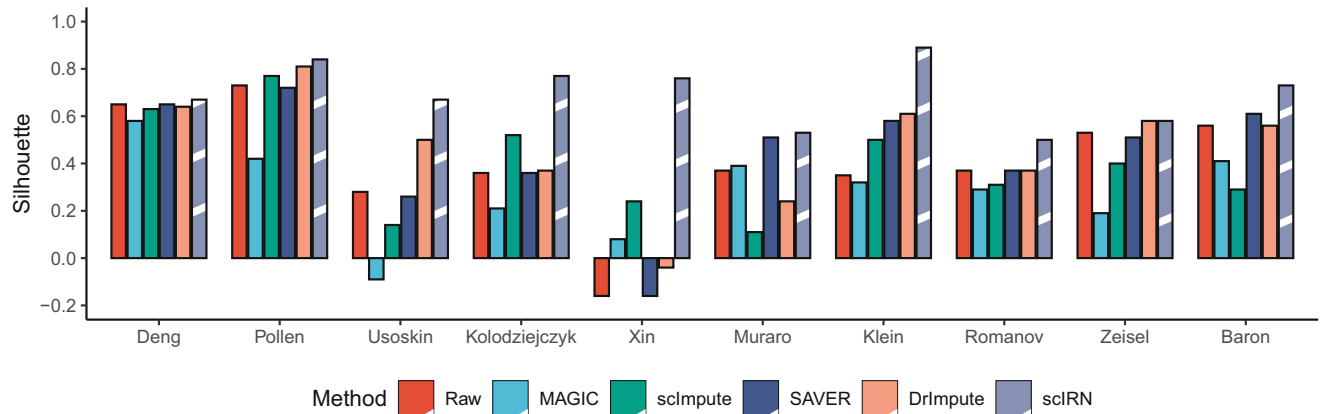


Fig. 52.5 Visualization quality using raw and imputed data, measured by silhouette index (SI). The *x*-axis shows the names of the datasets while the *y*-axis shows SI value of each method. scIRN outperforms other methods in all datasets

generate the transcriptome landscapes from raw and data imputed by MAGIC, scImpute, SAVER, DrImpute, and scIRN. We performed data visualization and calculated the silhouette index for each of the 10 datasets. Figure 52.5 shows the SI

values obtained for the raw data and data imputed by the five imputation methods. The figure shows that scIRN can improve the quality of data visualization in all datasets. scIRN also has the highest SI in each of these datasets. These results

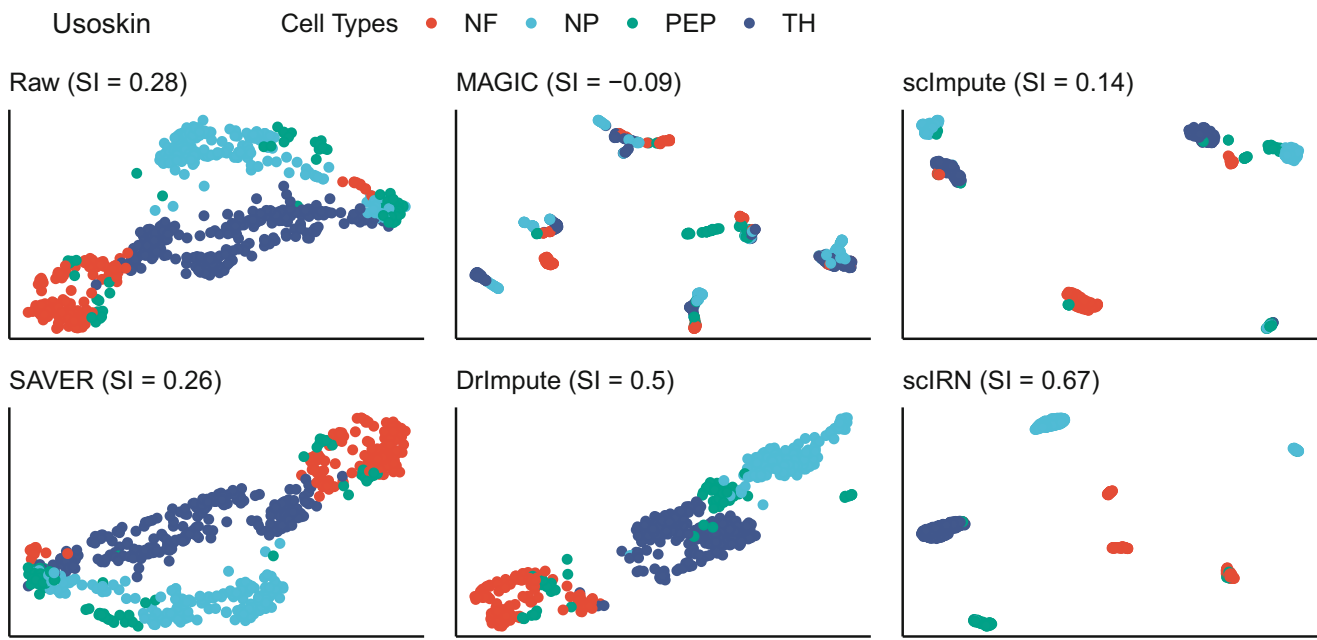


Fig. 52.6 Transcriptome landscape of the Usoskin dataset. The scatter plot shows the first two principal components calculated by UMAP. Different colors represent different cell types. The 2D representation

generated by scIRN has a clear structure, where cells from different groups are separated from one other

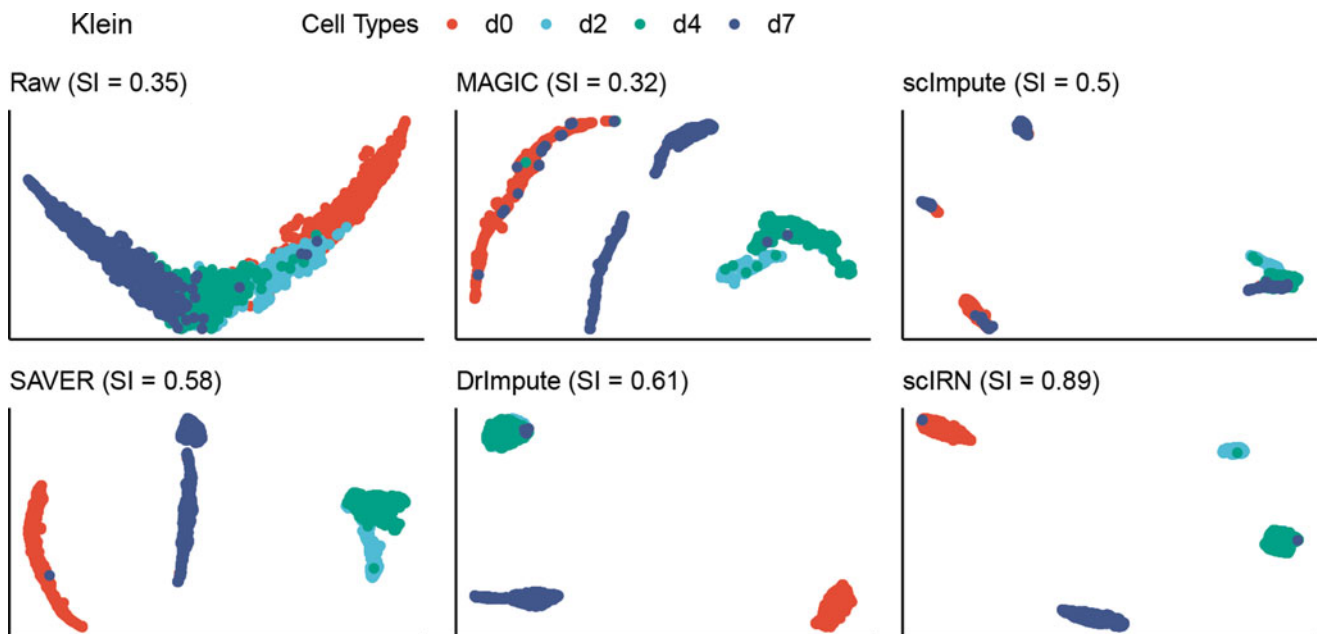


Fig. 52.7 Transcriptomics landscape of the Klein dataset. The scatter plot shows the first two principal components calculated by UMAP for

raw and imputed data. The 2D representation generated from scIRN has a clear structure, where cells from different groups are separate from each other

demonstrate that data imputation using scIRN would lead to a much better visualization of transcriptome landscapes compared to using raw data or data imputed by other methods.

Figure 52.6 shows the transcriptome landscapes of the Usoskin dataset. Using scIRN imputed data, UMAP was able to generate a clear representation, where cells from different groups are well-separated. When using data im-

puted by other methods, cells are usually mixed together. scIRN outperformed other imputation methods by having the highest SI value (0.67 compared to 0.28, -0.09, 0.14, 0.26, 0.5 of raw data, MAGIC, scImpute, SAVER, and DrImpute, respectively).

Figure 52.7 shows the transcriptome landscapes of the Klein dataset. The 2D representation of scIRN-imputed data

is the only one that has four separable groups, corresponding to the four real cell types. The landscapes generated using raw and data imputed by other methods have different cell types mixed together. The data imputed by scIRN has the highest SI value (0.89 compared to 0.61 of the second best).

52.4 Conclusion

In this article, we introduce a new method, scIRN, to recover the missing data caused by dropout events in scRNA-seq. We assess the performance of our approach using 10 single-cell datasets in a comparison with four current state-of-the-art imputation methods. Our analysis shows that scIRN outperforms existing approaches in improving the identification of cell sub-populations. scIRN also improves the quality of transcriptome landscapes generated by UMAP. A potential improvement of this research is to investigate the scalability of scIRN by analyzing datasets with higher number of cells. Another direction is to investigate the imputation method in other research applications, including pseudo-time trajectory inference and supervised learning. For future work, we will combine scIRN with current methods to improve the quality of downstream data analysis in the context of gene networks [40–47] and multi-omics integration [48–53].

Acknowledgments This work was partially supported by NASA under grant number NNX15AI02H (subaward no. 21-02), by NIH NIGMS under grant number GM103440, and by NSF under grant numbers 2001385 and 2019609. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

1. A.P. Patel, I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D.P. Cahill, B.V. Nahed, W.T. Curry, R.L. Martuza, D.N. Louis, O. Rozenblatt-Rosen, M.L. Suvà, A. Regev, B.E. Bernstein, Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**(6190), 1396–1401 (2014)
2. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**(6167), 193–196 (2014)
3. P.A. Darrah, J.J. Zeppa, P. Maiello, J.A. Hackney, M.H. Wadsworth, T.K. Hughes, S. Pokkali, P.A. Swanson, N.L. Grant, M.A. Rodgers, M. Kamath, C.M. Causgrove, D.J. Laddy, A. Bonavia, D. Casimiro, P.L. Lin, E. Klein, A.G. White, C.A. Scanga, A.K. Shalek, M. Roederer, J.L. Flynn, R.A. Seder, Prevention of tuberculosis in macaques after intravenous BCG immunization. *Nature* **577**(7788), 95–102 (2020)
4. L.D. Orozco, H.-H. Chen, C. Cox, K.J. Katschke Jr, R. Arceo, C. Espiritu, P. Caplazi, S.S. Nghiem, Y.-J. Chen, Z. Modrusan, A. Dressen, L.D. Goldstein, C. Clarke, T. Bhangale, B. Yaspan, M. Jeanne, M.J. Townsend, M.V.L. Campagne, J.A. Hackney, Integration of eQTL and a single-cell atlas in the human eye identifies causal genes for age-related macular degeneration. *Cell Reports* **30**(4), 1246–1259 (2020)
5. V. Kozareva, C. Martin, T. Osorno, S. Rudolph, C. Guo, C. Vanderburg, N.M. Nadaf, A. Regev, W. Regehr, E. Macosko, A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types. *bioRxiv* (2020)
6. P. Brennecke, S. Anders, J.K. Kim, A.A. Kolodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S.A. Teichmann, J.C. Marioni, M.G. Heisler, Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**(11), 1093–1095 (2013)
7. F. Buettner, K.N. Natarajan, F.P. Casale, V. Proserpio, A. Scialdone, F.J. Theis, S.A. Teichmann, J.C. Marioni, O. Stegle, Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**(2), 155–160 (2015)
8. P.V. Kharchenko, L. Silberstein, D.T. Scadden, Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**(7), 740–742 (2014)
9. S. Rizzetto, A.A. Eltahla, P. Lin, R. Bull, A.R. Lloyd, J.W. Ho, V. Venturi, F. Luciani, Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Sci. Rep.* **7**, 12781 (2017)
10. S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, I. Hellmann, The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016)
11. A. Haque, J. Engel, S.A. Teichmann, T. Lönnberg, A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**(1), 75 (2017)
12. W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, V. Shahrezaei, bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics* **36**(4), 1174–1181 (2020)
13. M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J.I. Murray, A. Raj, M. Li, N.R. Zhang, SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**(7), 539–542 (2018)
14. W.V. Li, J.J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018)
15. Z. Miao, J. Li, X. Zhang, scRecover: discriminating true and false zeros in single-cell RNA-seq data for imputation (2019). *bioRxiv*, p. 665323
16. W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, D.J. Garry, DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinf.* **19**, 220 (2018)
17. B. Tran, D. Tran, H. Nguyen, N.S. Vo, T. Nguyen, Ria: a novel regression-based imputation approach for single-cell RNA sequencing, in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)* (IEEE, New York, 2019), pp. 1–9
18. D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A.J. Carr, C. Burdzyak, K.R. Moon, C.L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**(3), 716–729 (2018)
19. F. Wagner, Y. Yan, I. Yanai, K-nearest neighbor smoothing for high-throughput single-cell RNA-seq data (2017). *BioRxiv*, p. 217737
20. A.M. Garay, E.M. Hashimoto, E.M. Ortega, V.H. Lachos, On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Comput. Stat. Data Anal.* **55**(3), 1304–1318 (2011)
21. Z.I. Botev, J.F. Grotowski, D.P. Kroese et al., Kernel density estimation via diffusion. *Ann. Stat.* **38**(5), 2916–2957 (2010)
22. A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2003)
23. S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**(1–2), 91–118 (2003)
24. V.Y. Kiselev, K. Kirschner, M.T. Schaub, T. Andrews, A. Yiu, T. Chandra, K.N. Natarajan, W. Reik, M. Barahona, A.R. Green,

- M. Hamberg, SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**(5), 483–486 (2017)
25. D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H.N. Luu, T. Nguyen, Fast and precise single-cell data analysis using hierarchical autoencoder (2019). bioRxiv, p. 799817
 26. D.P. Kingma, M. Welling, Auto-encoding variational Bayes (2013). arXiv: 1312.6114 [cs, stat]
 27. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778
 28. A.A. Pollen, T.J. Nowakowski, J. Shuga, X. Wang, A.A. Leyrat, J.H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D.W. Kemp, M. Wong, B. Clerkson, B.N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L.S. Weaver, A.P. May, R.C. Jones, M.A. Unger, A.R. Kriegstein, J.A.A. West, Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**(10), 1053–1058 (2014)
 29. D. Usoskin, A. Furlan, S. Islam, H. Abdo, P. Lönnerberg, D. Lou, J. Hjerling-Leffler, J. Haeggström, O. Kharchenko, P.V. Kharchenko, S. Linnarsson, P. Ernfors, Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18**(1), 145–153 (2015)
 30. A.A. Kolodziejczyk, J.K. Kim, J.C. Tsang, T. Illic, J. Henriksson, K.N. Natarajan, A.C. Tuck, X. Gao, M. Bühler, P. Liu, J.C. Marioni, S.A. Teichmann, Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**(4), 471–485 (2015)
 31. Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A.J. Murphy, G.D. Yancopoulos, C. Lin, J. Gromada, RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* **24**(4), 608–615 (2016)
 32. M.J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M.A. Engelse, F. Carlotti, E.J. de Koning, A. van Oudenaarden, A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**(4), 385–394.e3 (2016)
 33. A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D.A. Weitz, M.W. Kirschner, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**(5), 1187–1201 (2015)
 34. R.A. Romanov, A. Zeisel, J. Bakker, F. Girach, A. Helysaz, R. Tomer, A. Alpár, J. Mulder, F. Clotman, E. Keimpema, B. Hsueh, A.K. Crow, H. Martens, C. Schwindling, D. Calvigioni, J.S. Bains, Z. Máté, G. Szabó, Y. Yanagawa, M.-D. Zhang, A. Rendeiro, M. Farlik, M. Uhlén, P. Wulff, C. Bock, C. Broberger, K. Deisseroth, T. Hökfelt, S. Linnarsson, T.L. Horvath, T. Harkany, Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20**(2), 176–188 (2017)
 35. A. Zeisel, A.B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Bethsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, S. Linnarsson, Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**(6226), 1138–1142 (2015)
 36. M. Baron, A. Veres, S.L. Wolock, A.L. Faust, R. Gaujoux, A. Vetere, J.H. Ryu, B.K. Wagner, S.S. Shen-Orr, A.M. Klein, D.A. Melton, I. Yanai, A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**(4), 346–360 (2016)
 37. L. Hubert, P. Arabie, Comparing partitions. *J. Classif.* **2**(1), 193–218 (1985)
 38. E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I.W. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**(1), 38–44 (2019)
 39. P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
 40. H. Nguyen, D. Tran, B. Tran, B. Pehlivan, T. Nguyen, A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief. Bioinf.* (2020), bbaa190
 41. T. Nguyen, A. Shafi, T.-M. Nguyen, A.G. Schissler, S. Draghici, NBIA: a network-based integrative analysis framework-applied to pathway analysis. *Nat. Sci. Rep.* **10**, 4188 (2020)
 42. T.-M. Nguyen, A. Shafi, T. Nguyen, S. Draghici, Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.* **20**(1), 203 (2019)
 43. H. Nguyen, S. Shrestha, D. Tran, A. Shafi, S. Draghici, T. Nguyen, A comprehensive survey of tools and software for active subnetwork identification. *Front. Genet.* **10**, 155 (2019)
 44. T. Nguyen, C. Mitrea, S. Draghici, Network-based approaches for pathway level analysis. *Curr. Protoc. Bioinf.* **61**(1), 8–25 (2018)
 45. T. Nguyen, C. Mitrea, R. Tagett, S. Draghici, DANUBE: data-driven meta-ANalysis using UnBiased Empirical distributions—applied to biological pathway analysis. *Proc. IEEE* **105**(3), 496–515 (2017)
 46. T. Nguyen, D. Diaz, R. Tagett, S. Draghici, Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Nat. Sci. Rep.* **6**, 29251 (2016)
 47. T. Nguyen, R. Tagett, M. Donato, C. Mitrea, S. Draghici, A novel bi-level meta-analysis approach-applied to biological pathway analysis. *Bioinformatics* **32**(3), 409–416 (2016)
 48. H. Nguyen, S. Shrestha, S. Draghici, T. Nguyen, PINSPPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics* **35**(16), 2843–2846 (2019)
 49. A. Shafi, T. Nguyen, A. Peyvandipour, S. Draghici, GSMA: an approach to identify robust global and test gene signatures using meta-analysis. *Bioinformatics* **36**(2), 487–495 (2019)
 50. T. Nguyen, R. Tagett, D. Diaz, S. Draghici, A novel approach for data integration and disease subtyping. *Genome Res.* **27**(12), 2025–2039 (2017)
 51. A. Shafi, T. Nguyen, A. Peyvandipour, H. Nguyen, S. Draghici, A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Front. Genet.* **10**, 159 (2019)
 52. A. Shafi, C. Mitrea, T. Nguyen, S. Draghici, A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Brief. Bioinf.* **19**(5), 737–753 (2018)
 53. M. Menden, D. Wang, Y. Guan, M. Mason, B. Szalai, K. Bulusu, T. Yu, J. Kang, M. Jeon, R. Wolfinger, T. Nguyen, M. Zaslavskiy, S. Jang, Z. Ghazoui, M. Ahsen, R. Vogel, E. Neto, T. Norman, E. Tang, M. Garnett, G. Veroli, C. Zwaan, S. Fawell, G. Stolovitzky, J. Guinney, J. Dry, J. Saez-Rodriguez, Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* **10**(1), 2674 (2019)

Part X

Blockchain Technology

André Mury de Carvalho, Bruno Guazzelli Batista, and Adler Diniz de Souza

Abstract

IoT has been one of the emerging technologies for the past decade because of its facilities and it's also becoming widely used for monitoring purposes. Therefore, many questions are rising about the security of devices that can operate either in an everyday life or industry. It was found that IoT networks are using low-safety communication protocols that might be questionable and this paper reveals fair alternatives to replace these technologies using blockchain and smart contracts that will enforce the protection of these devices from security violation by setting up access control policies and roles to every peer in the network.

Keywords

Blockchain · Security · Internet of things · Network · Access control · Smart contract · Data immutability · Peer to peer · Distributed ledger · Data integrity

53.1 Introduction

Internet of Things (IoT) is one of the emerging technologies that promise an advancement in scientific and market scope. This happens because of the facilities that come with its utilization, from a simple presence sensor to systems that can monitor a huge machinery complex [11].

A. M. de Carvalho (✉)
 POSCOMP, Federal University of Itajuba, Itajubá, Brazil
 e-mail: mury@unifei.edu.br

B. G. Batista · A. D. de Souza
 Institute of Mathematics and Computing, Federal University of Itajuba, Itajubá, Brazil
 e-mail: brunoguazzelli@unifei.edu.br; adlerdiniz@unifei.edu.br

The applicability of this technology involves the insurance sector, smart environment, smart city, healthcare, industry 4.0 and a lot of others that can be developed using this concept [1]. Despite that, there are other major concerns in this area that surround security and access control issues [2].

Blockchain is a distributed database commonly associated with the public ledger concept, where all the digital transactions and events are shared among the network nodes. Each one of these transactions is verified by a majoritarian consensus method and when verified, they may be added into the network and can never be removed or modified. All the stored data is publicly verifiable by anyone who has access to the network, pushing reliability, integrity and verifiability to the network [1].

The Smart Contracts are a blockchain application to deal with the third party who intercepts a transaction between two (or more) subjects. These characters can be played by banks, registry offices, financial entities or any other [18] and became popular after the Bitcoin White Paper [12] and then, the Ethereum Network [20].

Considering all the characteristics described in the introduction, this paper tries to find interesting proposals to use the smart contract and blockchain concepts to ensure access control and security in an IoT network in the smart city context, in the point of view of network security and goals to evaluate the viability of this method to answer the following research questions (RQ):

53.1.1 Research Questions

1. **RQ1** According to the literature, is it possible to integrate IoT and a blockchain network?
2. **RQ2** How could smart contracts be used, in its ways, to protect data and integrity of an IoT network?
3. **RQ3** How feasible is it to secure an IoT network using blockchain and smart contracts?

53.2 Research Methodology

After the research and RQ definition, it was carried out a revision about the initial proposal followed by the research process definition from query strings, arbitrary search, inclusion and exclusion criteria, selection, quality evaluation and, lastly, data synthesis.

53.2.1 The Research Process

This process started in a pseudo-search for keywords about the overall theme using search managing tools to help the process of searching information and arbitrary search in academic databases such as IEEE, Scopus, Google Scholar, Academia.edu and CAFe.¹

The search term was firstly automatically suggested by Parsif.al² after the keywords and scope was defined and the first search string was:

```
S1=( "Blockchain" OR "Immutability"
OR "Peer To Peer" OR "IoT"
OR "Internet of Things"
OR "Remote Controllable Device"
OR "Sensor Network" OR "Smart Devices"
OR "Smart Contracts"
OR "Digital Contract"
OR "Distributed Ledger" OR "Ethereum")
AND ("Tool" OR "Technology" OR "Use")
AND ("Network" OR "Intranet"
OR "IoT Network")
)
```

As expected, S1 found a huge amount of papers because of the OR logical operators quantity that generalizes expressions selecting non-related results. S1 resulted in around 400.000 papers in Scopus.

Because of the infeasibility to analyze and summarize the papers, the term was refactored to minimize the false positive amounts, resulting in the following search string:

```
S2=((("Smart Contracts" OR "Blockchain")
,"Management"
) AND (
("Internet of Things"
OR "Smart Environment")
,"Smart Device"
) AND (
"Security","Integrity","Network"
)
)
```

S2 generated a feasible result of 35 papers in Scopus database therefore, CAFe's tool returned only one result, calling for a new modification in the search string:

```
S3=((("Smart Contracts" OR "Blockchain")
,"Management"
) AND (
"Internet of Things"
OR "Smart Environment"
) AND (
"Security","Integrity"
)
)
```

Looking at S1 and S2, it's easy to understand that the only difference is the word "Network" and, only because of this, a new search has resulted in 38 papers at CAFe. Repeated and non-related results were also discarded.

Following the idea of finding relevant papers to this SLR, it was also used the Google Scholar and Academia.edu platforms with arbitrary search keywords such as "Blockchain in IoT, IoT security management, Smart Contracts and IoT, Smart Contracts and Access Control management in IoT, Secure IoT access control management with blockchain, Trustability in IoT and Blockchain as IoT Infrastructure".

53.2.2 Inclusion and Exclusion Criteria

Some inclusion and exclusion criteria was used. As following, the inclusion criteria:

1. Papers, SLRs, books or surveys that are directly related to the theme;
2. Published in journals, conferences or magazines that are present in IT area; and
3. Written in English.

The following papers were excluded:

1. Includes keywords that was used in the search string but wasn't related to the theme;
2. Published before 2016 – except for Bitcoin white paper [12], Ethereum yellow paper [20] and SLR reference [9]; and
3. Similar or identical papers with the before selected.

53.2.3 Quality Evaluation

To evaluate papers, it was used the criteria discussed in subsection II-B and also a possible market relationship. As an example, [2, 11, 17] contain real market proposals, renewing the hope upon the subject of "Using Smart Contracts in IoT Access Control". In summary, it was selected a total of 24 papers shown in Tables 53.1 and 53.2.

¹CAFe is a brazilian aggregator similar to Scopus.

²<https://parsif.al>

Table 53.1 Chronological articles list [1]

Name	Description
Smart contracts [18]	Describes the fundamentals of smart contracts such as ideal concept and purposes of this paradigm.
Bitcoin: A peer-to-peer electronic cash system [12]	The Bitcoin White Paper is the main article related to blockchain concepts, describing the solved problem and how it was done.
Ethereum: A secure decentralised generalised transaction ledger [20]	The Ethereum yellow paper explains blockchain combining Szabo's smart contracts and Nakamoto's blockchain to build Ethereum virtual machine.
Blockchains and smart contracts for the internet of things [6]	This paper brings together ways to implement blockchain and smart contracts into the IoT
Blockchain technology: Beyond bitcoin [7]	Describes blockchain technology and some compelling specific applications in both financial and non-financial sector.
Towards a novel privacy-preserving access control model based on blockchain technology in IoT [13]	Describes how blockchain, the promising technology behind Bitcoin, can be very attractive to face those arising challenges.
The IoT electric business model: Using blockchain technology for the internet of things [23]	Proposes a redesign of c-business models with IoT and smart contracts
Blockchain-based structures for a secure and operate IoT [10]	Presents an application concept for a secure IoT operation using a blockchain based structure
Block chain technology [15]	Describes the whole blockchain concept showing all the processes and how them work to make blockchain a secure and unbreakable network.
A reference architecture for blockchain-based resource-intensive computations managed by smart contracts [8]	Addresses limitations of systems the performs resource intensive computations and manage private data using blockchain.
Blockchain and IoT integration: A systematic survey [14]	Analyzes the current research trends on the usage of Blockchain related approaches and technologies in an IoT context.
IoTChain: A blockchain security architecture for the internet of things [2]	Proposes an IoTChain combining OSCAR architecture and ACE authorization framework to provide an end-to-end solution to secure IoT access.

Table 53.2 Chronological articles list [2]

Name	Description
Do you need a blockchain? [21]	Analyses whether a blockchain is needed or not as a technical solution.
Blockchain design for trusted decentralized IoT networks [17]	Presents the opportunities and challenges of implementing blockchain and a use ease of integrating blockchain into an IoT framework.
Smart contract-based access control for the internet of things [24]	Investigates a critical issue in the IoT and proposes a smart contract based framework.
Decentralized IoT data management using blockchain and trusted execution environment [3]	Proposes a decentralized system of data management for IoT devices.
Multiple security certification system between blockchain based terminal and internet of things device: Implication for open innovation [5]	Addresses security threats in IoT and present schemes for developing a multi-security certification using blockchain.
Security issues on internet of things in smart cities [19]	Discusses a variety of issues related to internet of things in smart cities context.
SmartLock: Access control through smart contracts and smart property [22]	This paper presents a solution to rent properties using a smart contract to control door locking.
Towards secure IoT communication with smart contracts in a blockchain infrastructure [1]	Presents a Hyperledger fabric as a framework to secure IoT infrastructure with blockchain
A comprehensive survey on attacks, security issues and blockchain solutions for IoT and IIoT [16]	Discusses countermeasures proposed for the most relevant security threats in IoT
Using blockchain in IoT: Is it a smooth road ahead for real? [4]	Researches the challenges of blockchain in IoT to find out about its causes and cures.
Trust management in IoT enable healthcare system using Ethereum based smart contract [11]	The use of the blockchain technology to address security and privacy in IoT applications using Ethereum.

53.3 Results and Discussion

This section discusses results for each research question according to the references.

(i) *RQ1 – According to the literature, is it possible to integrate IoT and a blockchain network?*

As shown in [4], there are applications running that involves blockchain and IoT with a variety of goals among them. In relation to the Blockchain in IoT, the main goal is to improve the security, speed and coordination between devices and, when applied to hardware, secure data storage without using a third party.

There are many benefits to apply blockchain concepts into IoT [4] and because of the fact that the data can be directly shared between devices (P2P³) a several minimization of resource consumption become relevant to the industry.

In respect to [6], the authors studied the theme involving organizations that develop IoT and find that in the manufacturer point of view, the centralized architecture has a high maintenance level, considering the distribution of millions devices that deprecates fast. Otherwise, the consumer point of view shows a huge leak of trustability in these devices.

In an application environment [2, 3, 23], suggest infrastructure framework models to be applied in this concept. Particularly [3] proposes algorithms to secure data access using smart contracts to audit users and permissions without the need of a centralized system.

Another interesting application is the SmartLock described in [22]. It proposes a new way to control vacation renting such as AirBNB with a door locker that uses smart contracts to decide who can lock or unlock the door.

In this way, is possible to answer RQ1 positively. Is possible to integrate the blockchain into the IoT and also is possible to use smart contracts to enforce permission access.

(ii) *RQ2 – How could smart contracts be used to protect data and integrity of an IoT network?*

The blockchain is, by design, a secure infrastructure [12] and smart contracts are distributed applications – or scripts – stored in the blockchain and since they lay on blockchain, they have unique addresses and can be triggered when a transaction is fired using its address.

Each smart contract has their own independent state in the network, being able to take property of other assets in the network and also allows to transcribe business rules in code

format. This is important because any gap can be filled with a well programmed smart contract.

According to [24] there are a few access control models that are currently used:

1. RBAC – Role-based access control: restricts the network access based on the role of individual users in an organization allowing to access only the information the role permits;
2. ABAC – Attribute based access control: restricts the access based in policies that use arbitrary attributes to determine if the user is or not allowed; and
3. CapBAC – Capability based access control: uses tokens – or keys to authenticate users in the network and these users can share capabilities with other users using the principle of least privilege.

These access control models are based in a centralized architecture and have a unique point of failure, i.e., if the server fails, every node fail and to turn around this situation, the CapBAC model proposes a distributed structure where the devices send messages to each other to resolve the authentication request. Anyways, IoT devices tend to have a low storage capacity compromising this model and turning this model not fully trustable.

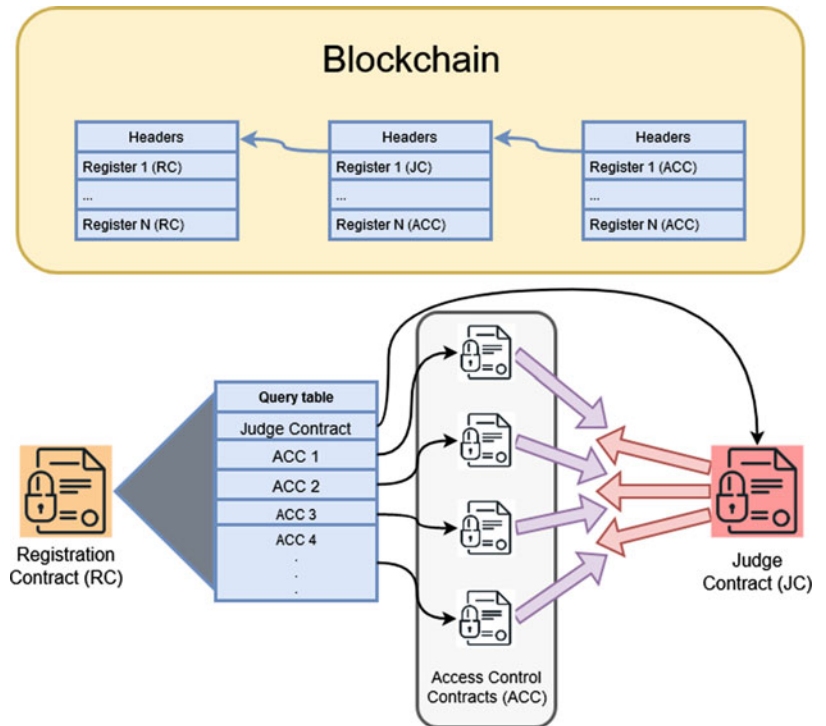
53.3.1 The Smart Contract System

The smart contract system described by [23] proposes a framework that consists of multiple access control contracts (ACC) and each one implements an access control for a pair of peers as illustrated by Fig. 53.1.

1. **Access Control Contract** is used to request control from a peer (object) to another (subject). It's assumed that a subject-object pair agrees to have multiple access control methods and each method is represented by one ACC. An ACC can associate to one and only one subject-object pair and the object requests to the subject. This method allows the ACC to implement both static and dynamic access control rights using predefined policies and checking the subject behavior.
2. **The Judge Contract (JC)** implements a misbehavior judging method, i.e., when any misbehavior is committed by a subject is reported by an ACC, the contract applies a corresponding penalty and can be based in the subject history so is also needed to keep a record of the misbehavior of all subjects. When the judgement is finished, the JC returns the decision to the ACC.
3. **The Register Contract (RC)** manages access control and misbehavior judging methods, maintaining a lookup table

³P2P: Peer-to-peer.

Fig. 53.1 Smart contract system [23]



registering all the needed information to execute all the methods.

The framework described by Zhang is far more complex and a deep explanation of its functions is not in this article scope but some important functions of this method are: call, register, update and remove access control methods and policies. Also, judge and AC contracts can be disabled if needed.

As a way to insert blockchain into an IoT context, [17] proposes an infrastructure model called Hyperledger. This model, illustrated by Fig. 53.2, is designed on a three layer architecture where each one of them realize a specific work:

1. Sensor layer connected to a Raspberry Pi responsible for generating a system I/O;
2. Edge layer that packages received data from sensors in a appropriate format to the ledger; and
3. Cloud layer where multiple pairs will be allocated to better failure resistance.

The hyperledger fabric [17] is a solution to pass through the network using the Docker Swarm tool – A tool supported by Docker to manage multiple containers hosted across multiple machines, much like Kubernetes – and the “Fabric” context is a permissioned blockchain where pairs of nodes do validation and the ordered nodes judge the consent and block the propagation to go on.

Integrating blockchain and IoT can bring a fast, decentralized and robust way to the management of a growing network

of devices, allowing a dynamic parameter configuration and assisting the auto administration. In context to RQ2 it is possible to conclude that smart contracts and blockchain may be used in an IoT network to perform as a middleware in the access control process with a secure storage, transparency, decision making and self maintenance.

(iii) *RQ3 – How viable is to secure an IoT network by a blockchain and smart contracts?*

With the conclusions of RQ1 and RQ2 is possible to bring some feasibility to this application of blockchain with a low opposition by manufacturers and users. IoT devices are up to 5 billion in 2018 and promise raising to 29 billion in 2022 and each one of these devices will interchange data with the internet and this situation will need better data management [14].

Kuzmin [10] presents an application concept of blockchain in a UAV⁴ network to control the airship traffic where the network is used to connect devices and store position and another data about all of them. This proposal has three parts:

1. 5G transmission towers to communicate devices, blockchain and users;
2. Devices to monitor; and
3. Controller devices such as laptops, tablets and others.

⁴Unmanned aerial vehicle.

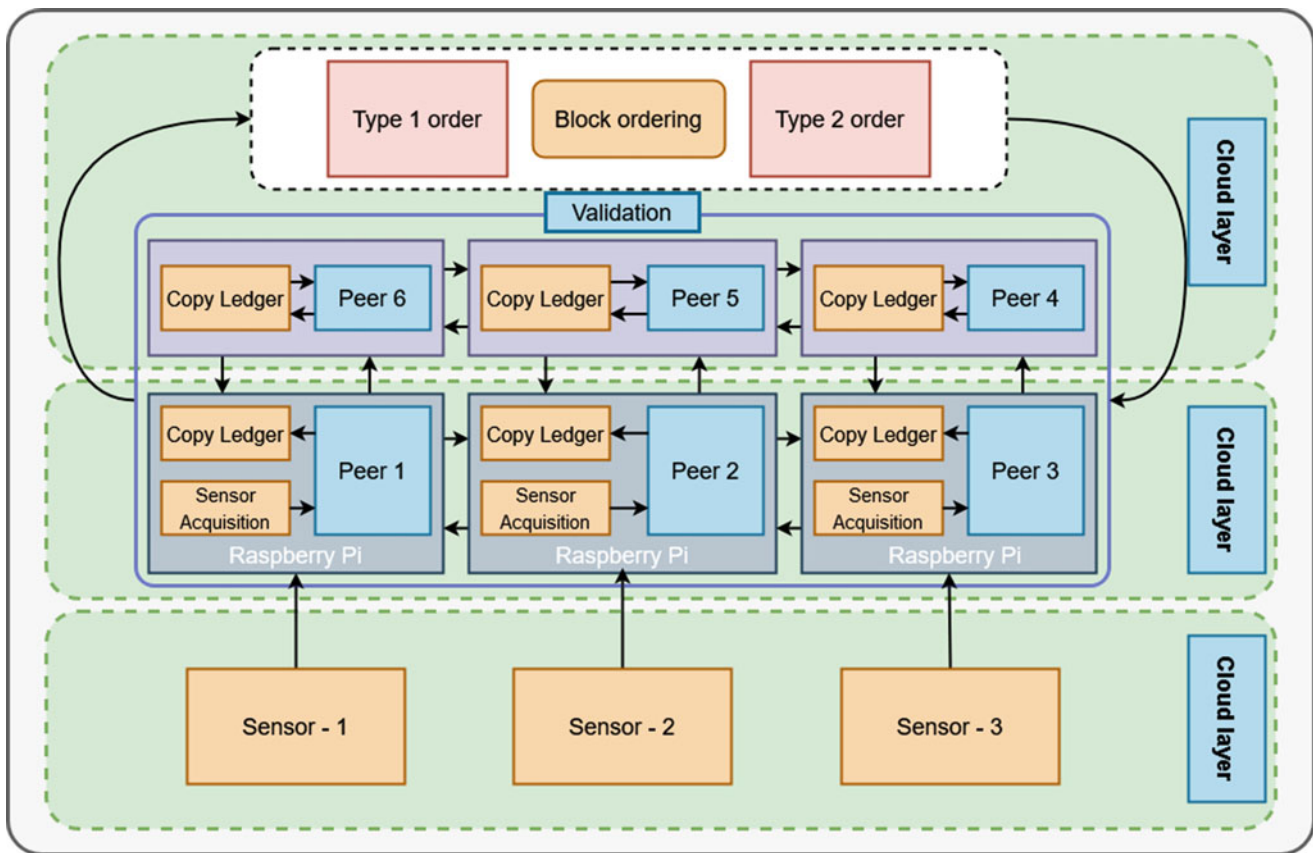


Fig. 53.2 Hyperledger fabric [17]

Indeed, some limitations around the use of an open blockchain network is the scalability. There are some steps that must occur from the beginning to the end of the transaction and it takes some time. For data reading the access is immediate, but the problem lays when the process is about storing data because of the mining process. To avoid this problem, it is possible to use a private permissioned blockchain (III-B), where the mining process is skipped [3].

In respect to the content gathered in this work, the answer to RQ3 may be positive. Although there are some limitations in a public blockchain scalability it is also possible to do a turnaround using a private permissioned blockchain and make a feasible way to integrate blockchain and smart contracts in an IoT environment. Using a system like blockchain to secure a network will make the environment safe and secure, ensuring that no malicious user will try to violate the environment causing undesired side effects.

53.3.2 to an enterprise network domain

A permissioned blockchain is a private network where all the users are known, similarly to an enterprise network domain to restrict the users who want access and secure the internal network. This way, the mining process that is responsible for

processing transactions in a public blockchain like Ethereum, doesn't need to exist because the users are already known and the transactions can be trusted [3].

According to [21] a permissioned blockchain allows only a predefined set of users to read and write assets in the network where there is a central entity to decide and give rights to each one of them. Also a permissioned blockchain is publicly verifiable meaning that any client is able to verify the veracity of a particular transaction by checking its transaction hash – or in a smart contract context, the contract itself.

As known, a public blockchain must have all of its transactions verified by the participants and the process of mining [6] is expensive, taking some time (around 10 minutes) that can't be wasted when an instant access is needed for example, when monitoring sensors in a factory or simply turning on a smart bulb.

53.4 Conclusion

Based on the material that was selected to write this paper, it was possible to clarify some questions about the theme "Smart Contracts in IoT access control" and answer the proposed research questions, allowing to highlight numerous

possibilities to the application in this kind of environment as an improvement to the current technology. IoT is an amazing way to make day life easier and deserves to be improved for better serving its purposes.

However some limitations related to the public blockchain were also noticed such as the scalability of connected devices issue because of the traffic amount in this public environment is directly proportional to the amount of connected devices and, there is a time period to be accomplished that is part of the blockchain design, but this problem can be solved with a private permissioned blockchain where the mining step can be skipped and only need the role or permission verification [3].

References

1. J. Ali et al., Towards secure IoT communication with smart contracts in a blockchain infrastructure. *arXiv preprint arXiv:2001.01837* (2020)
2. O. Alphand et al., IoTChain: A blockchain security architecture for the Internet of Things, in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, (IEEE, 2018), pp. 1–6
3. G. Ayoade et al., Decentralized IoT data management using blockchain and trusted execution environment, in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, (IEEE, 2018), pp. 15–22
4. S. Chandel, S. Zhang, H. Wu, Using blockchain in IoT: Is it a smooth road ahead for real? in *Future of Information and Communication Conference*, (Springer, 2020), pp. 159–171
5. B.-G. Choi, E.S. Jeong, S.-W. Kim, Multiple security certification system between blockchain based terminal and internet of things device: Implication for open innovation. *J. Open Innov. Technol. Mark. Complex.* **5**(4), 87 (2019)
6. K. Christidis, M. Devetsikiotis, Blockchains and smart contracts for the internet of things. *IEEE Access* **4**, 2292–2303 (2016)
7. M. Crosby et al., Blockchain technology: Beyond bitcoin. *Appl. Innov.* **2**(6–10), 71 (2016)
8. A. Kalogeropoulos, A reference architecture for blockchain based resource intensive computations managed by smart contracts. PhD thesis, Oct. 2018
9. S. Keele et al., *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical report, Ver. 2.3 EBSE Technical Report. (EBSE, 2007)
10. A. Kuzmin, Blockchain-based structures for a secure and operate IoT, in *2017 Internet of Things Business Models, Users, and Networks*, (IEEE, 2017), pp. 1–7
11. B.K. Mohanta, D. Jena, U. Satapathy, Trust management in IOT enable healthcare system using Ethereum based smart contract. *Int. J. Sci. Technol. Res.* **8**(9), 758–763 (2020)
12. S. Nakamoto et al., Bitcoin: A peer-to-peer electronic cash system (2008)
13. A. Ouaddah, A.A. Elkalam, A.A. Ouahman, Towards a novel privacy-preserving access control model based on blockchain technology in IoT, in *Europe and MENA Cooperation Advances in Information and Communication Technologies*, (Springer, 2017), pp. 523–533
14. A. Panarello et al., Blockchain and IoT integration: A systematic survey. *Sensors* **18**(8), 2575 (2018)
15. T. Prathyusha, M. Kavya, P. Sree Laxmi Akshita, Block chain technology (2018). <http://academicscience.co.in/admin/resources/project/paper/f20180328152225610.pdf>
16. J. Sengupta, S. Ruj, S.D. Bit, A comprehensive survey on attacks, security issues and blockchain solutions for IoT and IIoT. *J. Netw. Comput. Appl.* **149**, 102481 (2020)
17. J.C. Song et al., Blockchain design for trusted decentralized IoT networks, in *2018 13th Annual Conference on System of Systems Engineering (SoSE)*, (IEEE, 2018), pp. 169–174
18. N. Szabo, Smart contracts (1994). <https://goo.gl/geB8tm>
19. C. Thilagavathi et al., Security issues on internet of things in smart cities, in *Handbook of Research on Implementation and Deployment of IoT Projects in Smart Cities*, (IGI Global, 2019), pp. 149–164
20. G. Wood et al., Ethereum: A secure decentralised generalised transaction ledger, in *Ethereum Project Yellow Paper*, vol. 151 (2014), pp. 1–32
21. K. Wust, A. Gervais, Do you need a blockchain? in *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)*, (IEEE, 2018), pp. 45–54
22. M.X. Zapparoli, A.D. de Souza, A.H. de Oliveira Monteiro, Smart-Lock: Access control through smart contracts and smart property, in *16th International Conference on Information Technology-New Generations (ITNG 2019)*, (Springer, 2019), pp. 105–109
23. Z. Yu, J. Wen, The IoT electric business model: Using blockchain technology for the internet of things. *Peer Peer Netw. Appl.* **10**(4), 983–994 (2017)
24. Yuanyu Zhang et al., Smart contract-based access control for the internet of things. In: *IEEE Internet of Things Journal* **6**(2) (2018), pp. 1594–1605

Ibrahim Alkhamash and Waleed Halboob

Abstract

The Bitcoin technique faces several security challenges such as bitcoin wallet damage and attacks. This research presents a Bitcoin Wallet Security System (BWSS) to ensure the confidentiality and integrity of bitcoin wallet. The proposed system functions are distributed into two components and then implemented using Java programming language. A two-factor authentication is used to provide a more confidential bitcoin wallet while a backup mechanism is used to back up the wallet automatically or manually. Finally, the system is evaluated and refined after evaluation to come out with the required security features in a secure manner. The proposed system enables the bitcoin wallet owner to first protect his/her bitcoin wallet(s) with a second security layer using a two factor authentication and, second, back up his/her wallet in a secure and flexible ways. The system can be improved in the future to secure bitcoin wallets in mobile devices.

Keywords

Blockchain · Bitcoin · Wallet · Confidentiality · Integrity · 2FA · AES · Auditing · Backup · RSA

I. Alkhamash
Computer Science Department, Arab East Colleges for Graduate
Studies, Riyadh, Saudi Arabia

W. Halboob (✉)
Center of Excellence in Information Assurance, King Saud University,
Riyadh, Saudi Arabia
e-mail: wmohammed.c@ksu.edu.sa

54.1 Introduction

Bitcoin is a decentralized digital cryptocurrency. It provides local and international money transactions in a fast, and secure manner with low fees. The Bitcoin transactions are transferred from a person to another without going through government-based banking systems. When a sender sends a money, he adds a digital signature to the transaction message to inform that another user is now the owner of this Bitcoin and broadcasts it to what is called a bitcoin network. Finally, receiver of this message then sends the message to other machines to propagate the transaction information over the Internet [1].

With the Bitcoin, all transactions are known to any user holds the transactions database on his PC. The user can also verify all financial transactions by himself and do not have to rely on anyone for verifying the transactions integrity and authenticity [1, 2].

Unlike physical money, with Bitcoins there is no metal coins or paper money produced by government central banks. Even no government or central bank decides how much and when the Bitcoins are released to public. In fact, only 21 million bitcoins can be generated. The number of created bitcoins are increased by 50% every year. This means that people will still be able to create more Bitcoins until 2140. After that, no more Bitcoin can be created [3, 4].

A Bitcoin wallet is a file in a user's computer or phone. It saves public and private key pairs along with transactions relevant to this wallet. The public and private keys are used for sending and receiving Bitcoins. The private key is used to sign transaction messages and confirm the exchange of currency while the public key is distributed to payers to verify the transactions. User preferences are also kept in those wallet files that can and should be encrypted to mitigate the risk of losing the coins to a hacker [1].

Even if the Bitcoin system is secured using a complex cryptography, bitcoins worth millions of dollars have been stolen by attackers. The attackers just try to gain access to private keys of the victims. Basically, the private keys are found inside the Bitcoin Wallet which is just a file stored on the mobile phones or computers of the Bitcoin's owner or user. The number of malware used for targeting the Bitcoin has been increased as the Bitcoin use is increased [5].

This paper presents a bitcoin wallet security system (BWSS). The goal of the proposed system is to protect the Bitcoin wallet from stolen using malware or any kind of attacks. The system ensures the bitcoin wallets confidentiality through two security mechanisms namely two-factor authentication and encryption. First, the user has to authenticate himself using a password as well as one-time password (OTP) sent to his mobile or email. Second, the bitcoin wallets (which are files) are encrypted with an advanced encryption standard (AES) then the AES's key is encrypted with a RSA public key. For the bitcoin wallet availability, the proposed system ensures that by proposing a secure backup method. In other words, the user can backup his encrypted bitcoin wallets to another folder (located in another disks) or to a remote backup storage such as email or cloud computing storage. The innovation hub of this system lies in overcoming the aforementioned challenges by defining the security threats of the Bitcoin wallet according to its characteristics and security requirements.

The outline of this paper starts with related works in Sect. 54.2 followed by, in Sect. 54.3, introducing the proposed system architecture. The system implementation is discussed in Sect. 54.4 then the results are discussed in Sect. 54.5. Finally, the conclusion and future work are presented in Sect. 54.6.

54.2 Related Works

The Bitcoin technology faces several security and privacy challenges that can be classified into four categories namely: double spending, wallet attacks, network attacks, and mining attacks [6].

Within the double spending, the owner spends the same Bitcoin in more than one transaction. The basic solution for this issue is that the seller should wait for several confirmations before releasing the service to the user. This will make the double spend for the attacker harder but not possible.

The second Bitcoin Wallet security issue "Bitcoin attacks", the bitcoin wallet is stolen and reused. It can be stolen using computer attacks, for example. The Wallet damage is another bitcoin wallet security issue where the damage of the bitcoin can be intentionally or via a computer virus wallet since the wallet is just a file.

The third category of current Bitcoin attack types is the network attack. Here, the attacker targets the implementation

and design of the Bitcoin. He may target also the Bitcoin's peer-to-peer communication networking protocols. Several types of attacks are classified under this category such as Denial of Service (DoS) and Distributed Denial of Service (DDoS) attacks.

Regarding the mining attack which is the fourth Bitcoin category as discussed early. First, it is used for generating a Bitcoin by solving a complex math problem. Second, it is used for verifying the integrity and correctness of the Bitcoin transaction and here the miner can illegally gain more resources to mine more transactions and get more rewards from the mining pool manager.

This research focuses on the second Bitcoin security and privacy issue which is the Bitcoin Wallet attack. The Bitcoin wallet, which is a file holds public and private key pairs along with the address. The address is generated from the public key using hash techniques as discussed in Sect. 54.1.

The Bitcoin Wallets have two main security challenges, which are [6]:

First, the wallet can be stolen and used by the hacker. There is no way to recover the stolen private key since no other trusted third party can help on this issue.

Second, the Bitcoin wallet can be damaged by a Virus or intentionally. So, the owner requires a secure backup mechanism to restore his wallet.

According to the Krombholz et al. [6], around 22.5\$ of Bitcoin already losted from users because of these security attacks or even because of losing the public key or damaging the Wallet intentionally.

Several researches in the literature investigate the Bitcoin wallet security issues. Bamert et al. [7] propose BlueWallet, a hardware-based solution (token for the authorization of transactions). The BlueWallet can also be used as an electronic wallet to secure the Wallet. But, this device can be also stolen or damaged and, as a result, losing or damaging the Bitcoin Wallet.

In 2014, Goldfeder et al. [8] proposed a software solution for addressing several Bitcoin security issues with more focus on the Bitcoin Wallet security. They proposed a signature-based two factor authentication for securing the private key corresponding with the Bitcoin. This solution provides more a better security for protecting the private key from attacks but not from damage.

Gentilal [9] investigated the TrustZone technology and the Bitcoin protocol, and showed the Bitcoin wallet security could be improved using the TrustZone. However, the TrustZone technology is a complex solution and provides several security features. Its complexity makes it useless. Second, it is difficult to be used in mobiles devices. Finally, backup the wallet will require backing up a lot of information which make this solution less efficient as well.

Jareck et al. [10], proposed a password-based solution to secure the Bitcoin Wallet online. The password is hashed and then the hash value is used as a second layer secu-

ity. The research used this method to generate a more secure password from that one selected by the user. This system worked only online and could be attacked by brute forcing the used password. The user also may choose a weak password which makes this solution less secure than others.

54.3 The Proposed System Architecture

The architecture of the proposed system – called Bitcoin Wallet Security System (BWSS) is divided, as shown in Fig. 54.1, into the following components as the:

- Bitcoin Two-Factor Authenticator (B2FA): Used to authenticate the user using two-factor authentication mechanism.
- Bitcoin Key Manager (BKM): This is used for generating and changing the user keys namely AES and public/private keys.
- Bitcoin Wallet Encryptor (BWE): A java-based subsystem that is responsible for encrypting the bitcoin wallet.
- Bitcoin Wallet Store (BWS): This is a store for saving the encrypted bitcoin wallet.
- Bitcoin Wallet Backuper (BWB): Used for making a backup from the user bitcoin wallet. The system supports both automatically and annual backup. This is to ensure the availability security feature when the user lost his wallet.
- Bitcoin Wallet Auditor (BWA): This subsystem is used for auditing all activities happened on the user wallets. So, the user can review his financial activities and review them, interface that is mainly used for receiving the user subscription requests and dispatching them to the subscription engine.

The BKM works as the following:

- The user enters his password during the first time of using the system (or even during changing his password as will be discussed later on in this section). The user password is never transferred or saved, instead the user has to enter it again whenever the password is required. It is saved as a hash value as discussed in the next step.
- The user's password is hashed with SHA-2 hash method and then from the hash value an AES key is generated. The AES key is never saved but its hash value is saved.
- RSA public and private keys are generated where each key here is saved in a separated file.
- Since the private key is a secret key so it is encrypted with the user AES's key.

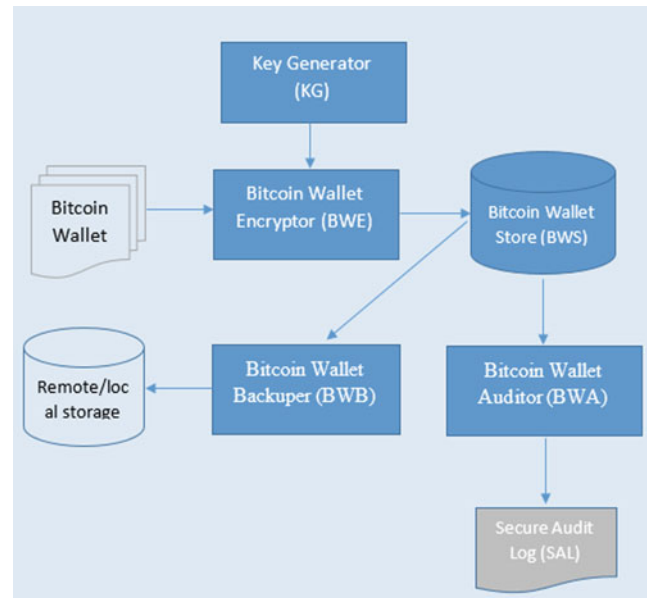


Fig. 54.1 The general architecture of the proposed system

Whenever the user wants to change his password, the following steps are executed:

- The user enters his old and new password.
- A new AES key is generated from the new password.
- The encrypted file that contains the private key is decrypted using the old AES key and then encrypted again with the new AES key.

The user public and private keys are generated based on RSA cryptosystems. Each key is a big prime integer number.

The BWS store used for storing the encrypted bitcoin wallet. It can be considered as a folder. All bitcoin wallet stored in this store are encrypted with AES and then the AES secret key is encrypted with the public key. It does not matter how many files (bitcoin wallet) are. So, the BWS is designed in such way it is a scalable component, means it can work with increased number of bitcoin wallet.

To ensure the availability of the bitcoin wallet, the BWB is used for making backup from the wallets and save it into a local and remote storage. The backup is made for all bitcoin wallets. The user can choose one of the following backup methods:

- Auto backup: In which, the bitcoin wallets are backed up into a pre-specified location and within a predefined period (e.g., daily, weekly, etc.).
- Manual backup: The user also can choose a manual back or whenever required. Using this backup method, the bitcoin wallet files are backed up to any location and at any time.

The secure audit log (SAL) component records an auditing information about the following:

- User financial activities: This includes user money transfer and receive.
- System activities: To record all processes activities such as changing password and keys as well as encrypting a wallet.

Figure 54.2 shows the flow chart of the proposed system. This flow chart shows how the bitcoin wallet is encrypted and then from time to time backed up. The execution steps of the systems are as follow:

1. The BKM component generates a private/public key for the user at the first start-up time and before encrypting any wallet.
2. The BWE receives the bitcoin wallet and encrypts them before storing them inside the BWS. The BWS is just a store for all encrypted bitcoin wallets.
3. The BWA audits all user financial and system processing activities for future use. All these activities are stored inside the Secure Audit Log (SAL). This log is secured by encrypting all of its contents are also encrypted using the user Key.
4. Finally, The BWB is used for making a backup from the user bitcoin wallets to a local or remote storage. The details of each component is discussed in the following sub-sections.

54.4 The System Implementation

The proposed solution is implemented in order to ensure that the designed system is workable. Therefore, it will be refined based on its implementation. The implementation also helps in measuring the proposed system to evaluate it. The proposed system is implemented using a Java programming language (JDK 1.8.0 version) with the assistance of several Java programming application program interfaces (APIs) and tools.

The first used tool is a NetBeans, which is a java integrated development environment (IDE). This IDE is used for devolving the proposed system to ensure that the system is a platform-independence. This means that the system can be executed in any operating system such as Windows, MacOS, Linux, and Solaris.

The Bitcoinj java package is used for working with the Bitcoin protocol and Bitcoin wallet as well as for managing the Bitcoin wallet and sending/receiving bitcoin transactions.

The Java Cryptography Extension (JCE) API is imported

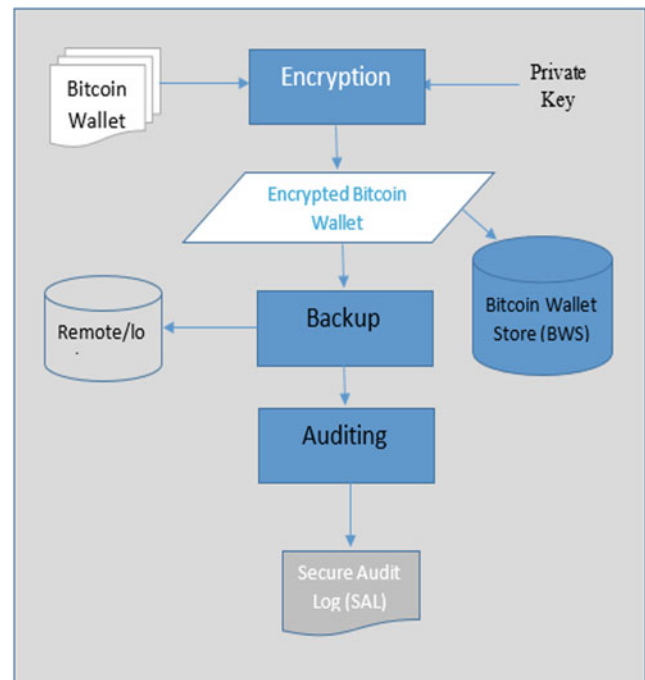


Fig. 54.2 The general flow chart of the proposed system

and used by the proposed system for random number generation, SHA-2 hashing method calling (to hash the user password), AES symmetric encryption for encrypting/decrypting the bitcoin wallets, and finally RSA asymmetric encryption to encrypt/decrypt the AES key. Using both AES and RSA in this way will provide a fast and secure encryption mechanism.

This is because the AES cryptosystem is fast but less secure while the RSA cryptosystem is slow but secure. Therefore, mixing both systems as described here will result a secure and fast encryption/decryption process.

The SMSLib java package is used for sending the SMS to the user. The SMS includes 6-digits Personal Information Number (PIN) with a limited lifetime (2 minutes).

Finally, the Log4j java package is used for implementing the SAL to record all user activities. This record will provide an accountability security feature.

54.5 Result and Discussion

The proposed system is evaluated using three main criteria, which are:

- Bitcoin wallet confidentiality.
- Bitcoin wallet availability.
- Bitcoin wallet accountability.

54.5.1 Bitcoin Wallet Confidentiality

For the system confidentiality, the proposed system ensures it through two security mechanisms, namely:

- Two factor authentication.
- Encryption.

The two factor authentication protect the system (in which the bitcoin wallet are managed encrypted) from any unauthorized access. The user has to authenticate himself first using his own password and, second, using a one-time password sent to his phone or email.

The password-based authentication here can be considered as the first line of defence in this system. The attacker cannot go forward without knowing the password. The user is recommended to select a strong password as well-known.

Since the password is not secure any more today in all systems. It can be attacked using a large number of attacking ways such as brute force, social engineering, etc. The proposed system also uses a second authentication factor, which is a one-time password (OTP). The user can choose to enable the OTP either using his phone number or email. In any way, the system will send the user 6-digits OTP with one-minute lifetime. This means that the OTP will expire with 60 seconds if the user didn't enter it. The user will need to ask for resending a new OTP.

Giving a user a chance to choose email-based or phone-based OTP increases the flexibility and availability of the system. The user may be not able to receive any SMS due to several reasons such as cost, roaming, and so on.

The two-factor authentication provides a very important feature in our system in terms of the incident response (when an attack is occurred). If an attacker gets the user password, he will not be able to access the bitcoin wallets directly as he needs also to enter the OTP. Also, the user will receive an SMS or email with a OTP and, this means, that at this point the user will know that his password has been attacked and the attacker is trying to access bitcoin wallet. The user can response to this incident (basically by changing his password) and before the full attack is occurred. The main goal of secure systems today is not to fully protect the system since providing a 100% secure system is not possible, but to detect any security incident (e.g., attack) before it is fully lunched or before the attacker gets an access to the information assets such as bitcoin wallet.

It can be concluded that the two-factor authentication used here is flexible and increases the availability of the system. Nevertheless, the main goal of using the two-factor authentication by this system is to prevent any unauthorized access to the system and, as a result, ensures the bitcoin wallet confidentiality as only the authenticated user can access the

system, in which the bitcoin wallets are stored, encrypted, and decrypted.

The second mechanism used by the proposed system is the encryption. If the attacker gains an access to the bitcoin wallets files by any attacking methods, the wallets' files are encrypted with a strong encryption process. In additions, each bitcoin wallets are encrypted with an AES key, which is then, is encrypted with the user RSA public key. The private key is always stored as a file and only decrypted at the runtime with the user AES's key. It is well-known that the RSA cryptosystem is the most secure cryptosystem. It can be observed that the proposed system ensures the bitcoin wallets confidentiality using different ways as presented in Table 54.1.

54.5.2 Bitcoin Wallet Availability

In cybersecurity, the availability means that data or service is available whenever required. The bitcoin wallet can be lost or damaged as discussed early. The proposed system ensures the bitcoin wallet availability through providing a bitcoin backup mechanism.

The user can choose two backup options, which are back up to another folder in another disk and backup to a remote storage (in a remote location) such as email account or cloud storage. There is no security risk while the user backs up his wallet to a remote storage operated or/and owned by a third party since the bitcoin wallet are encrypted as discussed in the previous section.

54.5.3 Bitcoin Wallet Accountability

To the best of our knowledge, the proposed system is the first bitcoin wallet security system that considers the accountability security feature. All user activities are recorded in a single and secure auditing log. The auditing log records several information, such as:

- Signing in date and time.
- Access bitcoin wallets with dates and times.
- Backed up bitcoin wallets with dates and times.
- Signing out dates and times.

It can be observed from the result discussed in this section that the proposed system also provides another security feature, which is a non-repudiation. The user cannot deny his actions if the system and bitcoin wallets are owned by an organization but managed by users. This security feature is provided through the using of two-factor authentication, strong encryption that uses a public/private key system, and finally auditing all user actions in a secure manner.

Table 54.1 Ensuring the bitcoin wallets confidentiality

Protection type	Security layers	Description
Authentication	Password	The first line of defense
	OTP	Helps in providing a second authentication layer as well as detecting any unauthorized access before the attacker gets a full access to the bitcoin wallet.
Encryption	RSA	Increases the encryption security by improving the encryption/decryption processes' complexity.

54.6 Conclusion and Future Scope

In this research paper, the security in bitcoin wallet is discussed and considered. A bitcoin wallet security system (BWSS) is proposed and developed using Java programming language with the assistance of other programming APIs and tools.

The proposed system ensures the confidentiality, availability and accountability of the bitcoin wallets using several security techniques namely two-factor authentication (password and OTP), two layers' encryption (AES and RSA), backup, and auditing log.

As a future work, there is a need for developing the same systems for mobiles as well as investigating the ability of developing and hosting this system in a web server to be provided as a service and used by a large number of user. To do so, this system must be refined to meet the requirements of the distributed environments.

References

1. R. Pallas, Bitcoin security. Master thesis, Tallinn University of Technology, 2012
2. A.B. Ayed, P. Taveras, T. BenYounes, Blockchain and IoT: A proposed security framework, in *17th International Conference on Information Technology–New Generations (ITNG 2020). Advances in Intelligent Systems and Computing*, ed. by S. Latifi, vol. 1134, (Springer, Cham, 2020). https://doi.org/10.1007/978-3-030-43020-7_17
3. D. Guegan, *The Digital World: 1 – Bitcoin: From History to Real Live* (2018)
4. C. Kugblenu, P. Vuorimaa, Decentralized reputation system on a permissioned blockchain for E-commerce reviews, in *17th International Conference on Information Technology–New Generations (ITNG 2020). Advances in Intelligent Systems and Computing*, ed. by S. Latifi, vol. 1134, (Springer, Cham, 2020). https://doi.org/10.1007/978-3-030-43020-7_24
5. M. Conti, S. Kumar, C. Lal, A survey on security and privacy issues of bitcoin. *IEEE Commun. Surv. Tutor.* **20**(4), 3416–3452 (2018)
6. K. Krombholz, A. Judmayer, M. Gusenbauer, E. Weippl, The other side of the coin: User experiences with bitcoin security and privacy, in *Financial Cryptography and Data Security: 20th International Conference, FC 2016, Christ Church, Barbados*, (Springer, Berlin, Heidelberg, 2017), pp. 555–580
7. T. Bamert, C. Decker, R. Wattenhofer, S. Welten, BlueWallet: The secure bitcoin wallet, in *International Workshop on Security and Trust Management* (2014), pp. 65–80
8. S. Goldfeder, J. Bonneau, E.W. Felten, J.A. Kroll, A. Narayanan, Securing bitcoin wallets via threshold signatures (2014). Available: <http://www.cs.princeton.edu/stevenag/bitcointhresholdsignatures.pdf>
9. M. Gentilal, P. Martins, L. Sousa, TrustZone-backed bitcoin wallet, in *Proceedings of the Fourth Workshop on Cryptography and Security in Computing Systems, ser. CS2 '17*, (ACM, New York, 2017), pp. 25–28
10. S. Jarecki, A. Kiayias, H. Krawczyk, J. Xu, Highly-efficient and composable password-protected secret sharing (or: How to protect your bitcoin wallet online), in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (March 2016), pp. 276–291



Ibrahim Alkhamash he obtained his bachelor degree in Computer Science from Taif University, and his master in Computer Science – Information Security Track from Arab Colleges in Riyadh – KSA. He is currently working as Cyber Security Manger and GM at the IT Department at Saudi Human Right Commission. He is a member of Cyber Security Committee and E-Government Committee. He had a wide range of professional certificates specialized in Projects Management, Business Development, Information Technology and Cyber Security.



Waleed Halboob Received his PhD degree in cybersecurity and digital forensics from University of Putra Malaysia (UPM), Malaysia, in 2015. His research interests include digital forensics, privacy, access control, trusted computing, etc. He has published more than 18 articles and is now serving as a reviewer for several international Journals and conferences.

Amjad Gawanmeh and Jamal N. Al-Karaki

Abstract

Disruptive technologies continuously and significantly alter the way people communicate and collaborate as well as the way industries operate today and in the future. To create new business models and opportunities, several combinations of disruptive technologies are being introduced nowadays. Among these technologies, cloud computing, IoT, Blockchain, artificial intelligence, social networks and media, big data, and 5G are mostly used. For instance, Blockchain technology made distributed solutions feasible and popular. On the other hand, big data and the social media are two contemporary technologies which lament significant impact on business and society. This paper presents a holistic approach to integration perspectives of these technologies considering many challenges like security and privacy. This paper also surveys the most relevant work in order to analyze how some of these technologies could potentially improve each other.

Keywords

Cloud computing · Social media · IoT · Blockchain · AI · Services · Next generation networks

A. Gawanmeh (✉)
 Department of Electrical Engineering, College of Engineering and IT,
 University of Dubai, Dubai, UAE
 e-mail: amjad.gawanmeh@ud.ac.ae

J. N. Al-Karaki
 Computer Information Science, Higher Colleges of Technology,
 Dubai, UAE
 e-mail: jkaraki@hct.ac.ae

55.1 Introduction

It is estimated that there are close to 8 billion connected IoT devices nowadays, a number that is expected to grow to 50 billion not far ahead. This will result in a huge IoT market that will drive the economy and continuously increase the interest and investment in these technologies, and hence accelerate their growth [1]. A new network of interconnected humans and devices, such as home appliances, gadgets, vehicles, and plant, has changed the way we deal with technology, and presented several new applications. It created several challenges related to data collection, storage, processing, analysis and communicate. The McKinsey Global Institute predicts that IoT could create economic impact between 2.7 trillion to 6.2 trillion annually by 2025 [2]. On the other hand, social networks and media are considered the main reason for the increased internet usage and for the generation of huge amounts of data [3].

With advancing of wireless and sensing technologies, Internet of Things applications have become more pervasive, cheaper, easier to implement, more practical, and hence more widespread. IoT includes physical devices with network functions, computational power, and connectivity functions [4]. In fact, several IoT applications were designed for particular use in the first place, however, with the widespread of several recent technologies, these have been migrated and adapted for everyday use. Practically, IoT is geared towards edge devices, this have created several challenges and opportunities as demonstrated in the literature, where other prevalent technologies, such as cloud computing, AI and big data significantly enhanced the performances and applications of IoT devices. For instance, this integration allowed reliable data collection, big data management, social data analytics, remote monitoring and control, and finally

machine to machine communications. Big data is generated from IoT infrastructure, this data requires aggregation, processing, storage, analysis, transmission, etc. This will require having some analytical, storage, and communication capabilities from edge nodes, hence, new infrastructures will be required to support this, which in turns creates several challenges and opportunities.

Blockchain offers a decentralized ability to authenticate and validate information exchange, it also provides necessary tools for verifying transactions through a group of unreliable actors using distributed, immutable, transparent, secure and auditable ledger. Figure 55.3 illustrates Blockchain operation and structure [5]. Artificial Intelligence (AI) is primarily used to develop methods that require an intelligent action and solve complex problems throughout self-learning. Integrating AI with IoT can provide powerful tools to solve several IoT issues using big data generated by IoT devices. AI also enables self-learning systems to provide solutions for various practical daily applications that ranges from smart grid to medical applications making use of the big data being generated.

Figure 55.1 shows new disruptive technologies [2] that can open several challenging opportunities, in particular for businesses, it is not easy to capitalize on these, since it is expected that iterations of such disruptive technologies will create new business models as they get more mature. As demonstrated in Fig. 55.2, the Hype Cycle for disruptive technologies grows very fast at the beginning, and it is driven by several collective factors and technologies [6, 7]. The evolutionary life cycle

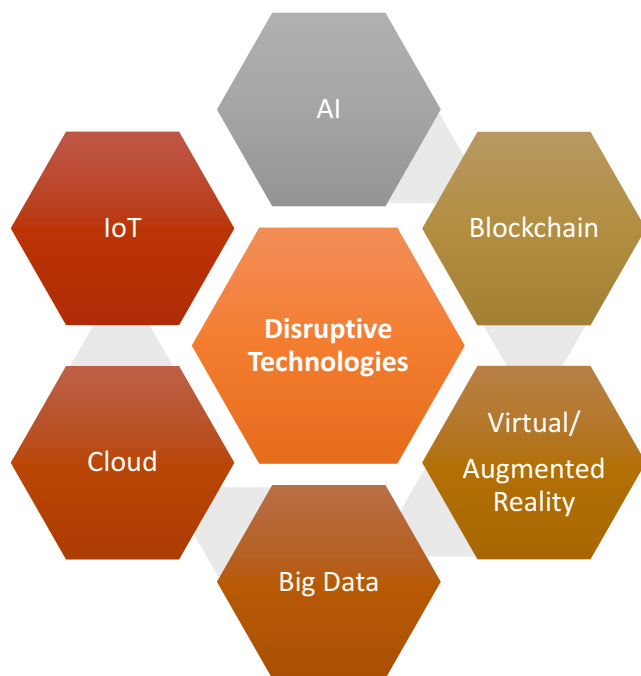


Fig. 55.1 Disruptive technologies

applies to technologies in their development and use. When a new high-technology emerges and gets popular, it not only challenges existing ones, but also forces them to integrate with it. This creates several innovations and opportunities. This paper provides a comprehensive analysis for challenges and opportunities that arise from the integration of disruptive technologies. This includes areas such as energy efficiency, governance, social profiling, tenancy and storage management, and several others (Fig. 55.3).

55.2 A Holistic Approach to Challenges in Disruptive Technologies

In this section, the overall interactions of disruptive technologies are discussed. We focus on security and privacy issues that result from such integration. First, we consider the intersection of Cloud, IoT and social media. To streamline our study, a holistic, flexible, and adaptive security framework needs to be established. Figure 55.4 shows a high-level sketch of integration and convergence of disruptive technologies and issues arise from this integration [8]. For example, IoT security is challenged with the low processing capabilities in the IoT devices, while IoT and social media applications depend entirely on the identity assurance as the source of data is fundamental in these technologies. In case of social media, users commonly share their sensitive information which could easily expose them to identity theft and stalking [9]. Identity theft is one of the major issues in social media, and needs to be addressed with sufficient evidence of identification and authorization of the entities involved E.

In IoT based systems, the entities are able to interchange data, such as; their identity, environmental information and physical properties, which is then used in the process of decision making. RFID an identification technology has supported the IoT concept for the identification of things in a unique manner. The management of identity in IoTs may raise possibilities of increasing security by putting together various authentication methods both for humans and machines.

Another key challenge in securing the future infrastructures is that of authentication and authorization. Several IoT devices may not have enough computational, storage and memory resources required to support security. In order to overcome these and other challenges, futuristic communications would require novel authentication systems constructed expending robust encryption and authentication systems. A similar approach is used by National Institute of Standards and Technology in choosing the compact SHA-3 as the innovative algorithm for “embedded” devices that have the capability to join electronic networks. The big data systems combined with analytics can provide opportunities for iden-

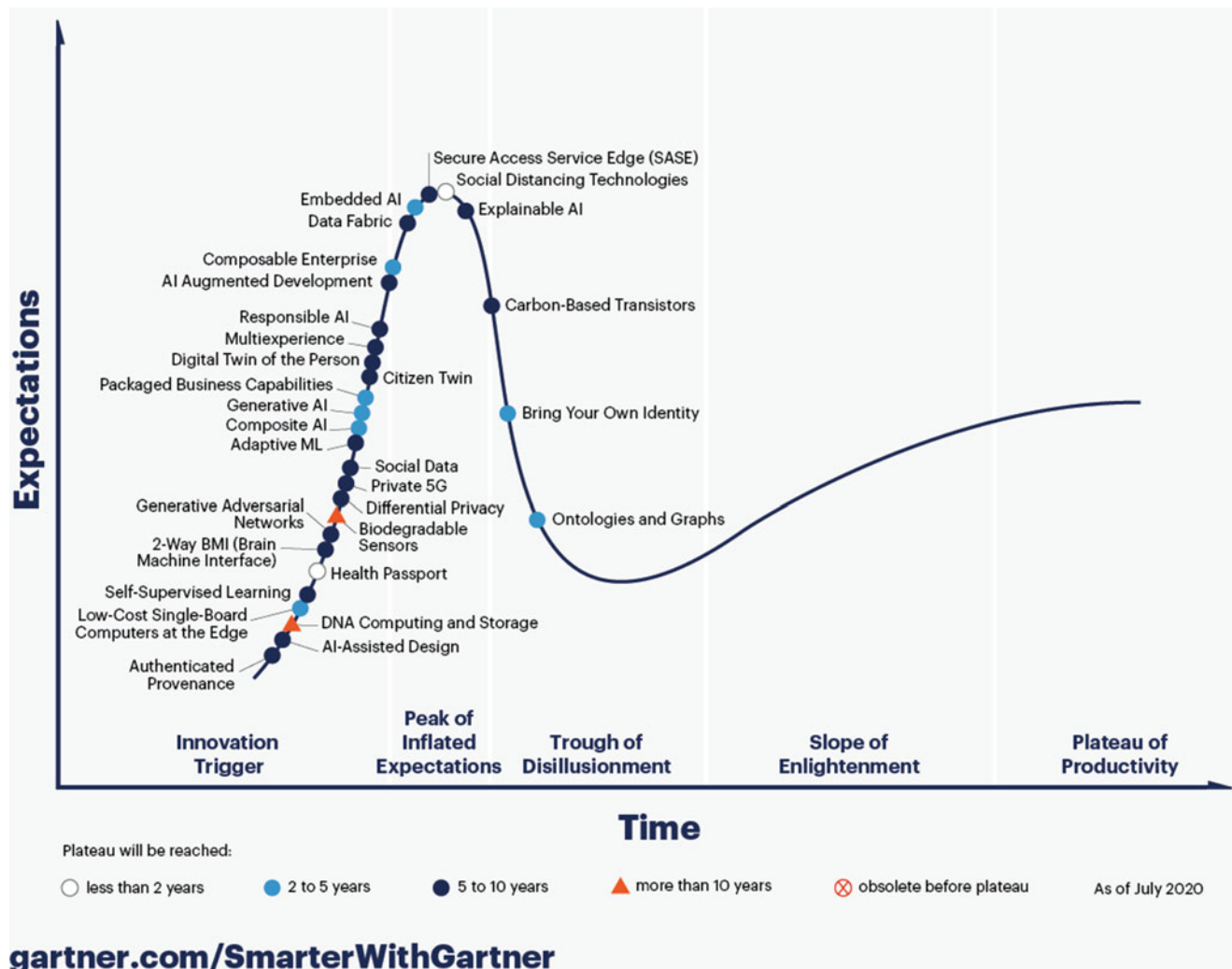


Fig. 55.2 Gartner Hype Cycle for disruptive technologies, 2020 [6]

tifying glitches in security related data by performing statistical analysis. Additionally, Obscurity and principal of least privilege, Strong passwords, Tough identities, and several others can also be included.

55.3 Intersection of Cloud, IoT and Blockchain

The unprecedented growth in new technologies such as IoT, cloud computing, and Blockchain has opened up new opportunities and provided new mechanisms to access and share information. The integration of these revolutionary technologies has proven to be invaluable. Blockchain, for instance can enrich both cloud computing and IoT through decentralized, reliable, trusted and traceable services. This can complement both cloud computing and IoT. Figure 55.5 demonstrates the challenges arise from integrating Blockchain with IoT [10].

One of the more unexpected use cases is blockchain working with IoT. Enterprises are starting to integrate IoT systems with distributed ledger technology. While IOTA, a distributed ledger designed to record and execute transactions between machines in the IoT ecosystem, has proven to be below expectations on many fronts, the use of distributed ledgers for storing data from IoT devices is already starting to happen. An additional incentive is the extra security these distributed ledgers provide. Although the key idea of Blockchain is simple, its implementation poses a several challenges such as storage, scalability, security, anonymity, data privacy, and legacy issues. Several countries are developing new laws in an attempt to regulate the use of virtual currencies. Blockchain technology is identified as the key to solve scalability, privacy, and reliability problems related to the IoT paradigm. It would improve IoT interconnection and the access of IoT data in Blockchain. On the other hand, the integration of Blockchain technology with the IoT is not trivial. For instance, Blockchain transactions requires the use of

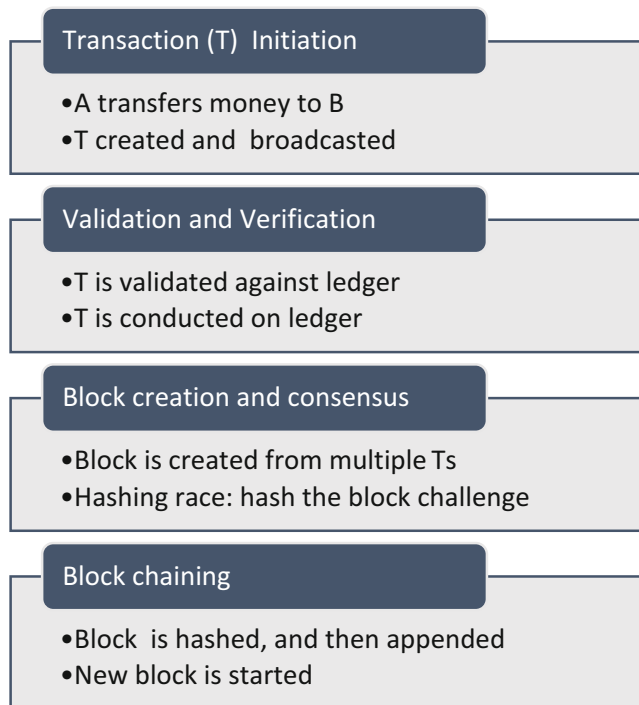


Fig. 55.3 The Operations of the Blockchain Technology

hashing algorithms to perform digital signature, and therefore IoT devices must be equipped with this functionality in order to be integrated with Blockchain.

55.4 Effect of Blockchain and AI on IoT Adoption

The AI, predictive analytics, IoT and Blockchain are all technologies that require strong data capture and use. Consequently, the way data is accessed will change to enable broader visibility and create cohesive ecosystems that support a convergence of data access and provides better operational and predictive capabilities. It may seem that AI, IoT, and Blockchain are three disparate corners of the technology world; however there is an obvious common thread which is data. Since data is the lifeblood of every enterprise and the pipe of business value across the various disruptive trends, various organizations should focus on the business value and data are often the key to such value.

Business leaders envision that IoT-driven organizations will start using AI to solve IoT-based problems. Industrial marketplace completely powered by Industrial IoT (IIoT) sensor fabric and deep learning algorithms will emerge soon. This will also facilitate a new class of deep learning algorithms that can detect actions. These algorithms will detect a person plus the action the person is taking. This “neural network” will just not identify a person in a video frame, it will also tag the person with an action such as, “person is

falling down.” This kind of application can heighten situational awareness of adverse safety events in real-time when a human misses it. This new kind of sensor will be a key driver behind IoT and digital transformation in manufacturing. Sensors with wireless connectivity, RFID tags and smart beacons are already gaining substantial traction in business. Manufacturers will be able to prevent delays, improve production performance, reduce equipment downtime and manage inventory.

The need for the integration of these technologies is particularly urgent now that algorithms used to extract insights from smart home devices to offer personalized experiences are a reality. Amazon’s Alexa is a major step forward. Further, many companies are already exploring Blockchain related use cases or implementing AI and IoT based environments across their business operations. While independently these technologies can bring great value to a company, greater business value can be obtained by combining these technologies together to effectively enable the “autonomous supply chain”. The future will also see greater convergence of these technologies, which will allow companies to then use deep performance insights to refine business processes, improve traceability of goods, and record and secure an archive of all digital interactions between companies and trading partners. However, in order for a company to maximize the benefits from an autonomous supply chain, it must have an end-to-end digital supply chain, a supply chain where every business transaction is exchanged electronically and not via paper.

Product innovation will be enabled by industry consolidation and technology integration. In fact, new solutions that will enable organizations to more easily integrate data across disparate systems are already starting to emerge and will lead to the creation of new services and changes in processes that save money and drive operational excellence. With the introduction of the long-promised 5G infrastructure, the IoT is finally ready for prime time. Home electronics will no longer need to be connected to the complexity of Wi-Fi setup and networks, the internal infrastructure provides its own network and that will enable more devices to be connected to the cloud. 5G sensors will reduce the cost of deployment and enable better connectivity and operational efficiency, in both enterprises and industry, where connectivity had once been difficult or impossible.

55.5 The Intersection of Clouds, IoT, 5G, and Social Media

In the IoT and 5G era several devices are being integrated within smart environment, whether at home, office, shopping centers, retail stores, or cars. Several other communication technologies such as RFID, NFC, WIFI, and smart sensors help to create networks that sense and communicate data and

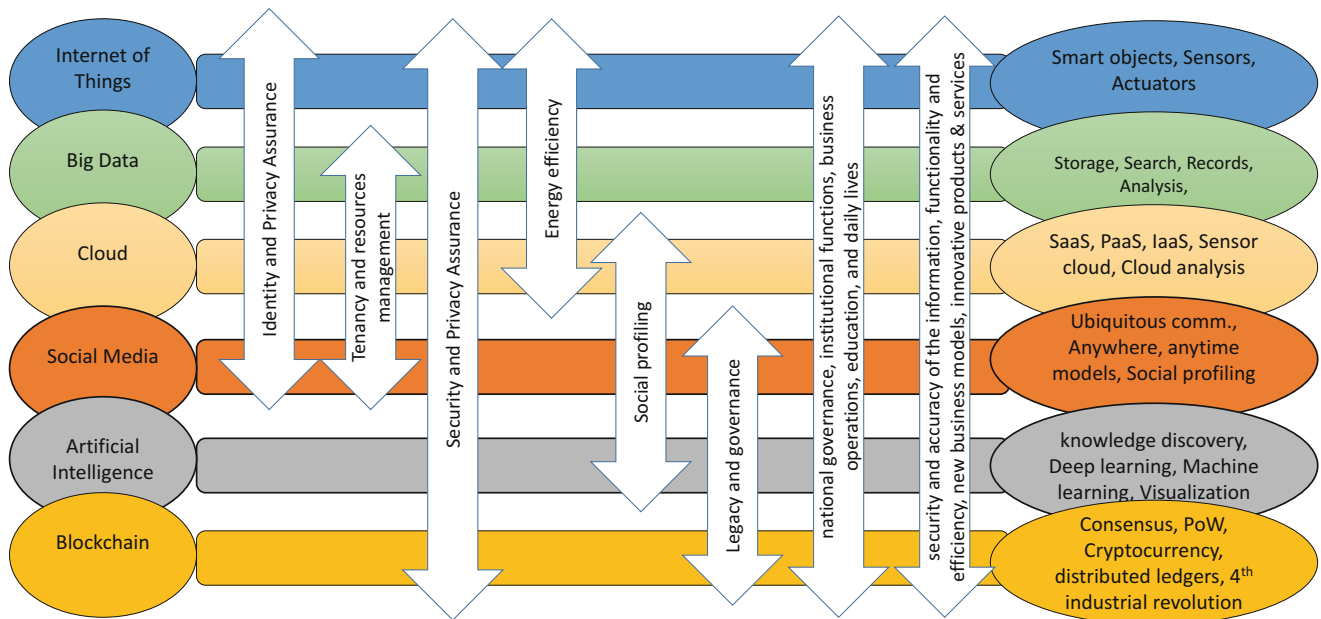


Fig. 55.4 Integration and convergence of disruptive technologies [8]

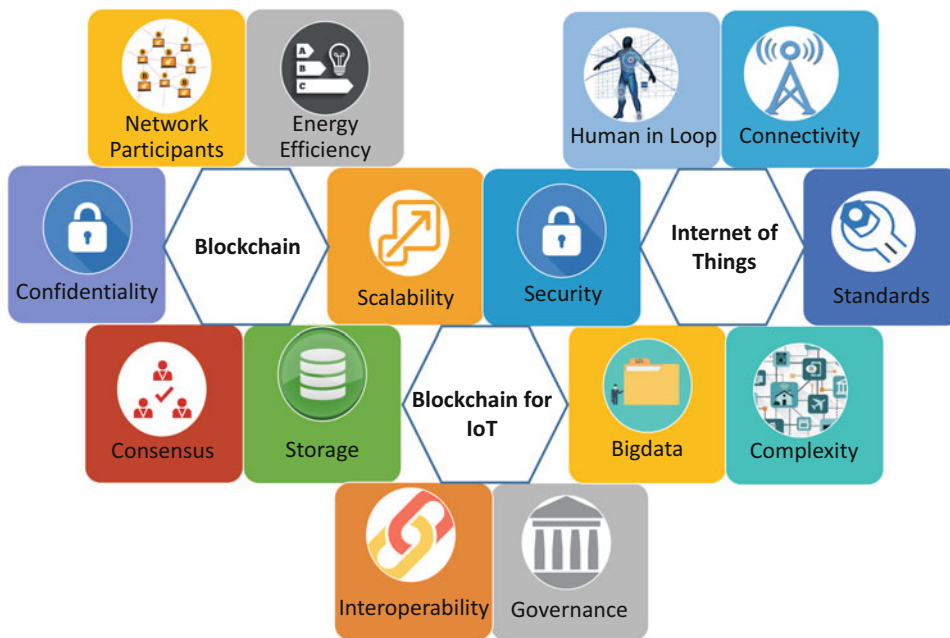


Fig. 55.5 Integration of blockchain and IoT

transparently. While several studies highlighted a number of potential applications for IoT [11, 12], it is still not clear on how cloud, IoT and social media can all be integrated and what are the challenges that may arise out of this. Table 55.2 provides a summary of application with highlights on where these technologies are being integrated.

These application, where IoT, cloud, or social media is used, will generate huge amounts of data. This data needs to be stored, filtered, analyzed, and probably presented and virtualized in smooth efficient and logical forms. In this sense,

IoT will play the role of data collection and aggregation, while cloud computing can provide virtual infrastructures that provide services such as visualization, analytics, storage, etc. On the other hand, social media provides necessary tools for timely and live communication between users, as well as service providers. Most of the platforms and solutions developed for AI and Blockchain are now hosted on the cloud. Integrating these technologies will require tremendous efforts, however, it will lead into systems with unforeseen capabilities.

55.5.1 Sensor Networks and Cloud Integration

Due to the limitations of Wireless Sensor Networks (WSNs), such as; memory, processing, energy, efficient management of data, communication, and scalability, recent research efforts are made for its improvement and as a result an architecture for sensor-cloud has been proposed. Recent study shows that sensor-cloud is currently becoming a point of focus for researchers in the field of WSN's and cloud computing [13–16]. In addition crowds sensing is another prominent domain that takes advantage of cloud, social networks as well as IoT [17–19]. By integrating WSN's and cloud computing the real-time processing and storage abilities are improved for the sensing data. Sensor-cloud provides several features such as dynamic provision analysis, resource optimization, collaboration, agility of services, quick response time [15]. Sensor-cloud Sensing-as-a-Service (SSaaS) is able to provide sensing data for more than one application simultaneously; thereby improving sensor resource utilization and sensor management. A sensor cloud can be the most tempting opportunity for a software as a service (SaaS) IoT. Any broadly based cloud provider could offer sensor cloud facilities, but the ISPs and telecommunication companies have the most buyer credibility in that space. SaaS for sensor cloud could be a launching point for other IoT cloud services. It would also drive competition and increase the adoption rate for IoT overall. It may be noted that in futuristic technologies cloud will not displace IoT, rather enhance it.

55.5.2 Sensors and Control Cloud Applications

The combination of sensor and control clouds resembles a management information base (MIB) commonly used to represent status and parameter control of routers and servers. Cloud applications could write to a variable used to change the state of the control. Network applications that read and manipulate MIB data could be used with the IoT.

55.5.3 Using Clouds for IoT and Big Data Analysis

The IoT analysis cloud is a set of services that correlates or analyses data to reach useful conclusions beyond digesting sensor data. For example, the IoT used on traffic and control signals for emergency vehicle movement would be ideal for finding the best route for ambulances or fire trucks, based on sensor data and available signal control points. Analysis can also be used to avoid revealing private information. A service could suggest a meeting point for friends without revealing their current locations.

55.5.4 Big Data and Social Media Integration

Both big data and the social media have significant impact on components of businesses, in particular, ones that deal with consumers. As social media has become a dominant tool that connects people, it has provided means to share information in many forms. This has resulted in the generation of huge amounts of data. In addition, it enables users to share their views, feeling, perception, satisfaction, and any other type of feedback. As a result, such data has provided a lot of new information about user trends. Social media has also enabled the collection of new types of data about behavior of people, this includes daily social habits, eating, sleeping, and hobbies. As a result, Big data and social media integration can have several consequences on any potential technology integration framework.

55.6 The Convergence of Blockchain, IoT, and AI

The disruptive technologies provide unprecedented opportunities to streamline and enhance existing processes, create entirely new business models, and develop innovative products and services for a new generation of consumers. It is expected that by 2022, every industry will implement some form of IoT technology. These IoT devices are gathering data at a massive rate, giving businesses near-infinite possibilities to improve their operational effectiveness. Current IoT conversations largely revolve around infrastructure (i.e. 5G networks and connected devices) while the focus should be more on how data can be processed into actionable insights using AI. The capabilities and limitations of syncing up IoT with AI is not well perceived, creating a key knowledge gap that, if bridged, could be the key to providing the most benefits.

While there is real feasibility of layering IoT, AI, and Blockchain to reap multiple benefits and create new opportunities, it is still not clear what challenges may arise from this. For one, Blockchain has the ability to create and build trust by offering transparency. Additionally, it facilitates faster transactions and reduces overall costs by cutting out the middlemen in various scenarios. Table 55.1 highlights the influence and need for different types of disruptive technologies for several contemporary applications. Below are three example of new business models in this contexts with other potential applications shown in Table 55.2:

- (1) **A retail powerhouse platform:** An AI-driven Blockchain can be efficiently used to collect real-time data from several sources such as retail, manufacturing, and consumer. Then, AI with machine learning capabilities process this big data in order to uncover useful patterns and archive knowledge that can be useful in many

Table 55.1 Disruptive technologies influence on different types of applications

	Innovative application	Cloud	IoT	Social media	5G	Big data	AI	Blockchain
1	Retail powerhouse platform	Low	High	Low	Medium	Medium	High	Medium
2	Service marketing	Medium	Low	High	Low	High	High	Medium
3	Customer support and engagement	Medium	Low	High	Low	High	High	Medium
4	Smart Cities	Medium	High	Medium	High	High	High	High
5	Health and Safety	High	High	Medium	Low	High	High	High
6	Digital 3D Printing	Low	High	Low	High	Medium	Medium	Low
7	Critical facility monitoring	Low	High	Medium	High	Medium	Medium	High
8	Supply chain	Medium	High	Low	High	High	Medium	High
9	Digital currency	Low	Low	Medium	Low	High	Low	High
10	Autonomous driving	Medium	High	Low	High	Medium	High	Medium
11	Electric cars charging	Low	Medium	Low	Low	Medium	Medium	Medium
12	Virtual augmented reality	Low	Low	Medium	Medium	Medium	High	Low
13	Crowd monitoring	Medium	Medium	High	High	High	High	Medium
14	Genomics applications	Low	Low	High	Low	High	High	Low
15	Pandemic Response	Medium	High	High	High	High	High	Medium
16	Mass vaccination	Medium	High	High	Low	High	Medium	Medium

directions. For instance, smart contracts provides more transparency about activities of retailers, which in turn increases trust and raises standards.

- (2) **Service, marketing, and customer engagement:** This can provide new tools for the interaction. Chatbots client on Facebook Messenger platform is a good example. A new paradigm will be able to do more tasks in terms of customer support and satisfaction given that they become more seamless in their communication capabilities. Finally, such tools will be able to make good use of AI and machine learning in order to be able to perform complex tasks. For instance, Chatbots will be able to answer complex questions and process and solve difficult requests from prospects and customers with the help of AI.
- (3) **Smart Cities with IoT, Blockchain, and AI:** The concept of a smart city cannot be achieved without proper paradigm. For instance, reducing resource consumption is key point in a smart city can. This cannot be achieved without using IoT enabled devices, big data analysis, machine learning, and cloud computing support. For instance, we need to optimize the electricity and water consumption; reduce traffic congestion and air pollution, enhance public transportation, etc. This requires real-time data collection and analysis, can be achieved through harnessing the power of big data and ubiquitous IoT devices with the integration of cloud computing and AI.

55.7 Conclusion

The integration of new technologies, such as IoT, AI, Blockchain, 5G, and social networking, enables a lot of challenging applications, including supply chain manage-

Table 55.2 Cloud, IoT, and social media applications

	Application	Cloud	IoT	Social media
1	Healthcare [20, 21]	x	✓	x
2	Transportation industry [22, 23]	✓	✓	x
3	Infrastructure monitoring [24, 25]	✓	✓	x
4	Environment monitoring [26]	✓	x	x
5	Smart cities [27, 28]	✓	✓	✓
6	Manufacturing industry [29]	✓	x	x
7	Water and electricity [30, 31]	✓	✓	x
8	Agriculture and farming [32, 33]	x	✓	x
9	Emergency events [34, 35]	✓	✓	✓

ment, healthcare, smart systems, industry 4 applications, and several others. Hence, it is still not clear what capabilities will be there when practical systems can make use of all of these technologies altogether. In this paper, we address most relevant challenges and opportunities that arise from the integration of these technologies. In addition, this paper has highlighted how several new innovative challenges are influenced by these new disruptive technologies.

References

1. S. Nižetić, P. Šolić, D. López-de-Ipiña González-de, L. Patrono et al., Internet of things (IoT): opportunities, issues and challenges towards a smart and sustainable future. *J. Clean. Prod.* **274**, 122877 (2020)
2. J. Manyika, M. Chui, J. Bughin, R. Dobbs, P. Bisson, A. Marrs, *Disruptive Technologies: Advances that will Transform Life, Business, and the Global Economy*, vol. 180 (McKinsey Global Institute, San Francisco, CA, 2013)
3. T. Kolajo, O. Daramola, A. Adebiyi, Big data stream analysis: a systematic literature review. *J. Big Data* **6**(1), 47 (2019)

4. G. Aceto, V. Persico, A. Pescapé, Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. *J. Ind. Inf. Integr.* **18**, 100129 (2020)
5. Z. Zheng, S. Xie, H. Dai, X. Chen, H. Wang, An overview of blockchain technology: architecture, consensus, and future trends, in *2017 IEEE International Congress on Big Data (BigData Congress)* (IEEE, New York, 2017), pp. 557–564
6. Gartner LLC, The Gartner Hype Cycle for Emerging Technologies. <https://www.gartner.com/>. Accessed 12 December 2020
7. K. Panetta, *Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017*, Gartner, Stamford (2017)
8. J.N. Al-Karaki, A. Gawanmeh, Security and privacy challenges of integrated disruptive technologies, in *2019 2nd International Conference on Signal Processing and Information Security (ICSPIS)* (IEEE, New York, 2019), pp. 1–4
9. G. Guo, Y. Zhu, R. Yu, W. Cheng-Chung Chu, D. Ma, A privacy-preserving framework with self-governance and permission delegation in online social networks. *IEEE Access* **8**, 157116–157129 (2020)
10. D. Pavithran, K. Shaalan, J.N. Al-Karaki, A. Gawanmeh, Towards building a blockchain framework for IoT. *Cluster Computing* **23**(09), 1–15 (2020)
11. P. Matta, B. Pant, Internet of things: genesis, challenges and applications. *J. Eng. Sci. Technol.* **14**(3), 1717–1750 (2019)
12. S. Balaji, K. Nathani, R. Santhakumar, IoT technology, applications and challenges: a contemporary survey. *Wirel. Person. Commun.* **108**(1), 363–388 (2019)
13. S.K. Madria, Sensor cloud: a cloud of sensor networks, in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)* (IEEE, New York, 2017), pp. 2660–2661
14. M. Fazio, A. Puliafito, Cloud4sens: a cloud-based architecture for sensor controlling and monitoring. *IEEE Commun. Mag.* **53**(3), 41–47 (2015)
15. T. Wang, H. Luo, W. Jia, A. Liu, M. Xie, MTES: an intelligent trust evaluation scheme in sensor-cloud-enabled industrial internet of things. *IEEE Trans. Ind. Inf.* **16**(3), 2054–2062 (2019)
16. T. Ojha, S. Misra, N.S. Raghuvanshi, H. Poddar, DVSP: dynamic virtual sensor provisioning in sensor–cloud-based internet of things. *IEEE Internet Things J.* **6**(3), 5265–5272 (2019)
17. T. Luo, J. Huang, S.S. Kanhere, J. Zhang, S.K. Das, Improving IoT data quality in mobile crowd sensing: a cross validation approach. *IEEE Internet Things J.* **6**(3), 5651–5664 (2019)
18. V. Gelardi, J. Godard, D. Paleressompoulle, N. Claidière, A. Barrat, Measuring social networks in primates: wearable sensors versus direct observations. *Proc. R. Soc. A* **476**(2236), 20190737 (2020)
19. P. Bellavista, D. Belli, S. Chessa, L. Foschini, A social-driven edge computing architecture for mobile crowd sensing management. *IEEE Commun. Mag.* **57**(4), 68–73 (2019)
20. S. Beitelspacher, M. Mubashir, K.M. Beshar, M.Z. Ali, Prioritizing health care data traffic in a congested IoT cloud network, in *2020 IEEE Wireless Communications and Networking Conference Workshops* (IEEE, New York, 2020), pp. 1–6
21. J.N. Al-Karaki, A. Gawanmeh, M. Ayache, A. Mashaleh, DASS-CARE: a decentralized, accessible, scalable, and secure healthcare framework using blockchain, in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)* (IEEE, New York, 2019), pp. 330–335
22. T.M. Anand, K. Banupriya, M. Deebika, A. Anusiya, Intelligent transportation systems using iot service for vehicular data cloud. *Int. J. Innov. Res. Sci. Technol.* **2**(2), 80–86 (2015)
23. G. Xu, M. Li, L. Luo, C.-H. Chen, G.Q. Huang, Cloud-based fleet management for prefabrication transportation. *Enterprise Inf. Syst.* **13**(1), 87–106 (2019)
24. Z. Lv, B. Hu, H. Lv, Infrastructure monitoring and operation for smart cities based on IoT system. *IEEE Trans. Ind. Inf.* **16**(3), 1957–1962 (2019)
25. D. Anderson, T. Gkountouvas, M. Meng, K. Birman, A. Bose, C. Hauser, E. Litvinov, X. Luo, Q. Zhang, Gridcloud: infrastructure for cloud-based wide area monitoring of bulk electric power grids. *IEEE Trans. Smart Grid* **10**(2), 2170–2179 (2018)
26. S. Corbellini, E.D. Francia, S. Grassini, L. Iannucci, L. Lombardo, M. Parvis, Cloud based sensor network for environmental monitoring. *Measurement* **118**, 354–361 (2018)
27. M.S. Kamal, S. Parvin, K. Saleem, H. Al-Hamadi, A. Gawanmeh, Efficient low cost supervisory system for internet of things enabled smart home, in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)* (IEEE, New York, 2017), pp. 864–869
28. A.U. Rehman, R.A. Naqvi, A. Rehman, A. Paul, M. Tariq Sadiq, D. Hussain, A trustworthy snot aware mechanism as an enabler for citizen services in smart cities. *Electronics* **9**(6), 918 (2020)
29. Y. Liu, X. Xu, Industry 4.0 and cloud manufacturing: a comparative analysis. *J. Manuf. Sci. Eng.* **139**(3), 034701-1–8 (2017)
30. J. Lloret, J. Tomas, A. Canovas, L. Parra, An integrated IoT architecture for smart metering. *IEEE Commun. Mag.* **54**(12), 50–57 (2016)
31. J.L. Stănică, G. Căruțașu, A. Pirjan, C. Coculescu, IoT cloud solution for efficient electricity consumption. *J. Inf. Syst. Oper. Manage.* **12**(1), 45–57 (2018)
32. S. Parvin, S. Venkatraman, T. de Souza-Daw, K. Fahd, J. Jackson, S. Kaspi, N. Cooley, K. Saleem, A. Gawanmeh, Smart food security system using iot and big data analytics. in *16th International Conference on Information Technology-New Generations* (Springer, New York, 2019), pp. 253–258
33. N.M. Imran, M. Won, Reducing operation cost of lpwan roadside sensors using cross technology communication. Preprint. arXiv: 2009.12471 (2020)
34. Z. Xu, Y. Liu, N. Yen, L. Mei, X. Luo, X. Wei, C. Hu, Crowdsourcing based description of urban emergency events using social media big data. *IEEE Trans. Cloud Comput.* **99**, 1 (2016)
35. C. Arbib, D. Arcelli, J. Dugdale, M. Moghaddam, H. Muccini, Real-time emergency response through performant IoT architectures, in *International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, May 2019, Valencia (2019)

Esraa Dbabseh and Radwan Tahboub

Abstract

The Automatic Meter Reader (AMR) for energy consumption is one of the most important issues in smart cities, as meters and electricity companies suffer from insecurity. The Internet of Things (IoT) can be used to achieve effective and reliable AMR in real time. Blockchain is a very advanced technology and technology that can be used to secure transactions such as meter readings and meter control. It is based on the idea of sequencing data blocks in a secure and distributed manner. Ensures the security of the blockchain representing the data in each block (reading the meter for example). The block is created and verified by many devices distributed on a network. Blockchain can be implemented in various ways and environments such as the Ethereum platform. In this paper, we will present a new automated meter reading platform using blockchain technology to meet the complete security requirements of AMR systems. When DoS attack was launched, requests did not affect the data itself, but the response speed of the blockchain network to incoming transactions from the servers was reduced very slightly. In addition, the results show that Blockchain could provide a promising technology that can participate in securing network meters.

Keywords

AMR · Blockchain · Ethereum · IoT · Smart contract

E. Dbabseh (✉)
College of Graduate Studies and Scientific Research, Palestine
Polytechnic University, Hebron, Palestine
e-mail: 131076@ppu.edu.ps

R. Tahboub
Departement of CE, College of Information Technology and Computer
Engineering, Palestine Polytechnic University, Hebron, Palestine
e-mail: radwant@ppu.edu

56.1 Introduction

The development of technology and the existence of the Internet play a vital role in all areas of life. One of the most important of these applications and technologies is the Automatic Meter Reading Technology (AMR). The automatic meter reading is intended for remoting, monitoring and controlling of the local energy meters. This technology enables regular meter reading without visiting every home. The reading can be achieved with a micro-controller that continuously monitors and records the meter reading in the databases [5]. The data is transmitted by using the internet to achieve efficient and reliable AMR in real time. The automatic meter reading will be the consumer friend, because it takes care of all the problems that a consumer could potentially present and fully in control of the power board. Automated metering uses online applications, these apps mostly deal with databases that must be protected which they often contain sensitive data. Database serves a large number of users. Attention needs to be focused on many attacks, whether single or group. Therefore, the applications should provide security and don't allow unauthorized persons to access and read this data without knowing its details; Otherwise, the data loses its meaning and value [8]. There has been a lot of results aimed at sending data over the Internet of Things and storing it in the cloud and other solutions. But these solutions are still central and there is a third party controlling them [3]. One of the technologies that has appeared recently is blockchain technology that has changed a lot of applications, because it has a lot of features that makes it stands out from the rest. The most important of these features is that it is distributed and decentralized. The idea of the blockchain is a network that contains a large number of users who sends transactions to each other. These transactions must be validated and all values verified before entering the blockchain. Data and transactions are accepted

and added if it's validity is proved by so-called miners. Validate these transactions by solving mathematical equations that are present within each transaction. Thus, makes this technology powerful and distinct from others [14].

56.2 Background

This chapter provides a background and an introduction to the automatic meter reading and its importance, after which an overview of the blockchain and its implementation methods will be presented, and then an overview of the ethereum smart contracts.

AMR technology allows the automatic collection of consumption and diagnostic data from meter. This can lead to decrease labor costs and more accurate billing. Basically, the meters are equipped with sensors that register the meter automatically. These sensors then transfer the data electronically. The data is then transferred to a central database for billing and analysis, which reduces the chance of input errors. These meters help in accessing accurate and up-to-date data from meters and closely monitor and control energy costs. These meters also help solve many problems, especially in reaching remote areas that are difficult to reach periodically [6].

Cryptographic technologies are based on the science of cryptography, which protects sensitive information that cryptographic techniques use to restrict behavior instead of using trusted third parties. Blockchain is a series of continuous data records called blocks that are linked together and secured with cryptography. Based on a peer-to-peer topology, the blockchain is a distributed ledger technology that allows data to be stored globally on thousands of servers while allowing anyone on the network to view anyone else's entries in nearly real time. This makes it difficult for a single user to control with the network. A blockchain begins when the user sends transaction to other user and transactions are enabled, although they can be tracked, but are anonymous. Public keys are cryptographically generated addresses stored in the blockchain. The amazing feature of blockchain is that public keys are never associated with a real-world identity. The blockchain is a decentralized network that contains a shared, unchanging authority ledger, which means that the information inside the authority leader is open and available to all. The blockchain network is transparent and this feature makes its reliable [9]. This technology, is an easy way to pass information in a completely secure way. The block is created and verified by thousands of devices distributed on the network. It is added to the previous group of chains and stored across the network. If someone wants to forge a single transaction, they are forced to change the entire previous chain and this is almost impossible. So the data is not changeable and the database is managed using a peer-to-peer network. Ethereum is a decentralized computing plat-

form. It generates a cryptocurrency token known as Ether. Ethereum was released in 2015 and is open source software for major chains used in cryptocurrencies and ether. It enables the creation and operation of smart contracts and distributed applications (DApps) without any interruption, fraud, control or interference from a third party. Ethereum helps developers create and deploy distributed applications, not only because it is a platform but also a complete Turing programming language. There are two types of accounts: the first type is externally owned accounts that are controlled by private keys, and the second type is the contract account that is controlled by its contract code. Miners use this algorithm to verify the validity of a transaction before adding it to the chain of blockchain [11].

56.3 Literature Review

There are many different new technologies used to read meters, the most important one is that which uses Internet technology. AMR can be classified into wired and wireless systems, depending on the medium used to transmit the readings. Both systems have advantages and disadvantages. Measuring power over the wire is an expensive system because it requires a change in the infrastructure compared to wireless units. WIFI is more suitable for this type of application because WIFI has become one of the common facilities everywhere [1]. Here we will talk about different techniques and methods described by the authors. Wessam Mesbah, Senior Member [7] Authors discuss a new way to secure meter reading against tampering or malfunction to discover and correct customer attacks that aimed to change smart meter readings. The idea of linear error-correcting block codes was used, which was used in a system of Communications to detect and correct errors in data transmission. It was suggested to use codes with some modifications in order to detect false readings in some meters that measure electrical energy. Xingyuan Fan, Chun Zhou, Ying Sun, Jinyang Du and Ying Zhao [4] Authors proposed designing a new generation of meter-based meter reading system through Narrow Band Internet of Things (NB-IoT) technology. The direct connection between the power meter and the main station system was achieved. The intermediate equipment measurement station was deleted and the complexity of the current intensive copy platform structure was reduced. The central management and real-time monitoring capacity of the power meter was improved, and the group coverage rate was improved effectively. Also the number of connections to one base station can be increased easily. The smart meter connects directly to the NB-IoT network via a remote connection unit that supports the standard NB-IoT connection. Pranav Singhal, Sakshi Upadhyay, Sheenam and Annurudh Pratap Upadhyay [10] Authors proposed the smart power meter

uses IoT technology to monitor and manage energy. The system was designed to resort in a local server and database when the internet connection was resumed. In this research, a digital power meter consisting of blinking LED signal was used. It is coupled with a controller using Optocoupler. This microcontroller reads data and sends it to the IoT platform using the WiFi unit. Sequentially transmits data to the IoT platform for display where power meter readings can be accessed globally. The reading of energy consumed is displayed on the platform website. Energy consumption reports are generated daily and can be monitored around the clock at any time.

56.4 Blockchain Based AMR Framework Design

The proposed model aims to create a hybrid model consisting of a cluster of servers with a blockchain, in order to maintain electricity meters and read data, as well as many of the features available. This improves the safety of multiple attacks and also monitors and controls meters. The proposed system has two main components. The first is a network of smart meters distributed in the regions of customers. It consists of a control panel through which data is read and displayed on the LCD screen. The second component is the blockchain network, the data coming from the meters will be used as input to the blockchain network, whether for examination, control, or preservation. This app is used to store meter reading for all users. These components will be used to prevent multiple security attacks. Figure 56.1 shows an overview of the frame structure. As shown in the figure, any server must participate in the blockchain network in order to preserve the network and its data as well as obtain the benefits of using this framework. Each group of meters belongs to a specific server. The use of blockchain within the proposed framework is to preserve data and prevent multiple attacks.

The server is part of a blockchain network. One of the server's tasks is to send the received data from the meters and save it to the blockchain network. Each server stores the data in the blockchain network through transactions. In a smart contract, the addresses of servers that are allowed to read and write in the blockchain network are defined. Since the server is part of the blockchain network, each server has a unique address in the blockchain network. If the address is one of the allowed to write on the blockchain network so the data will be received from the server and stored. To increase safety of the received data from the meter to the server, it will write the data after making sure that the current reading value for this meter is greater than or equal to the previous value, in order to ensure the correctness of the upcoming data.

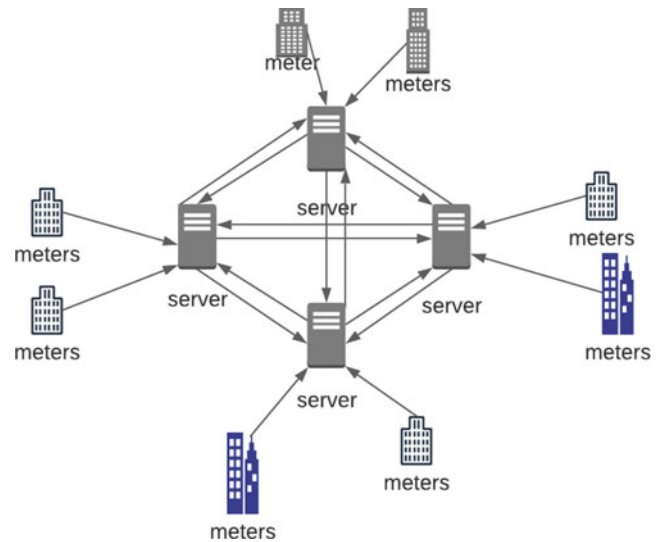


Fig. 56.1 General architecture of the proposed framework

A smart contract defines the rules between different organizations in executable code. Applications call a smart contract to create the transactions that are recorded in the ledger. Using the blockchain network, we can turn these contracts into actionable software. Smart contracts are applications that are published in the authority book and executed independently as part of the validation of transactions. Once the contract is created, the address and balance are also created for it. Smart contract is written in several languages, but the language used to write nodes is solidity, which is the JavaScript language developed for writing smart contract. Figure 56.2 shows, the smart contract that was used in the proposed work is a deal within the Ethereum blockchain. In this decade, addresses of users able to write into the blockchain network were identified. For example, in our research the server is part of the blockchain that receive data from the meters. The servers are users blockchain who can send the data to the smart contract and which has to save the data.

56.5 Proposed Framework for Blockchain Based AMR

In our proposed work, meter reading was secured between servers nodes in the blockchain. When an attacker attempts to write in a blockchain, the address of the server that sent the data to the network will be verified from the addresses allowed to write to the blockchain network. Likewise, if the attacker was able to know a valid address, he wouldn't also be able to write, because he does not have the private key of the server on which the transaction is signed. So this network will be resistant to this attack [12].

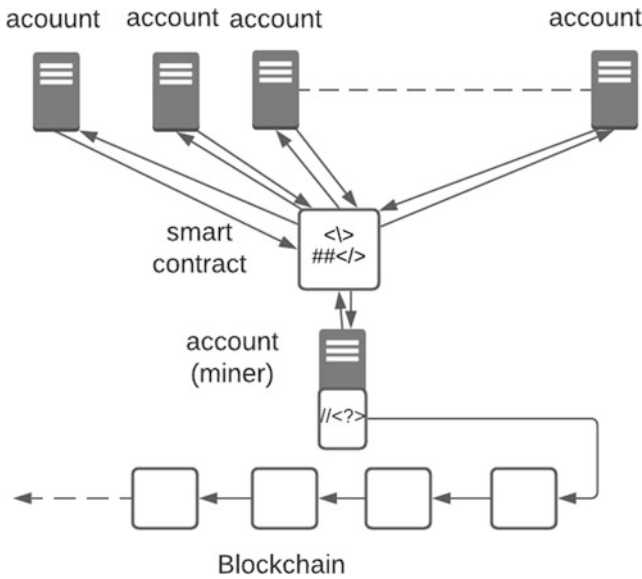


Fig. 56.2 Smart contracts architecture

Since the blockchain systems are independent and do not depend on any technology, the availability in the meter network using blockchains will be ensured. For the denial of service attack to succeed, requests are sent to one node because the system is central. As the blockchain is a decentralized system linked to multiple nodes, a denial of service attack needs access to the different nodes at the same time to damage the network. This prevents denial of service attacks through the decentralization of the network [13].

Replay attack resends a previously sent message. And the data that are sent is correct, but duplicate. In a blockchain network when the attacker tries to send a previous reading saved in the blockchain network, this technology prevents this attack because every transaction (i.e. reading a meter) has an infrequent timestamp is used. It is placed so that a distinction is made between the original (first) transaction and that no repeat transaction is accepted after it [2].

56.6 Implementation of the Framework

This section explains the practical approach and the implementation of the blockchain network on the prototypes used. Also, it demonstrates the effectiveness of the proposed framework and discuss response time and Latency on Transaction. Latency is the time taken to transmit a packet over the network. The main concerns of this research are writing on the blockchain network and reading from that network when needed. Whenever an attempt is made to implement writing or to make a connection to a blockchain network to read through specific addresses that are permitted to interact

Table 56.1 Response writing data to the proposed smart contract

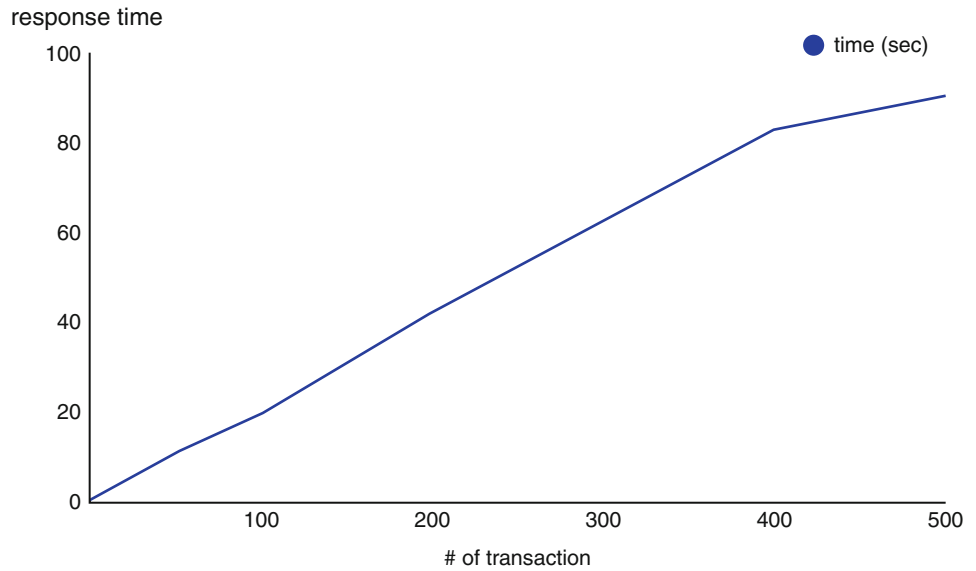
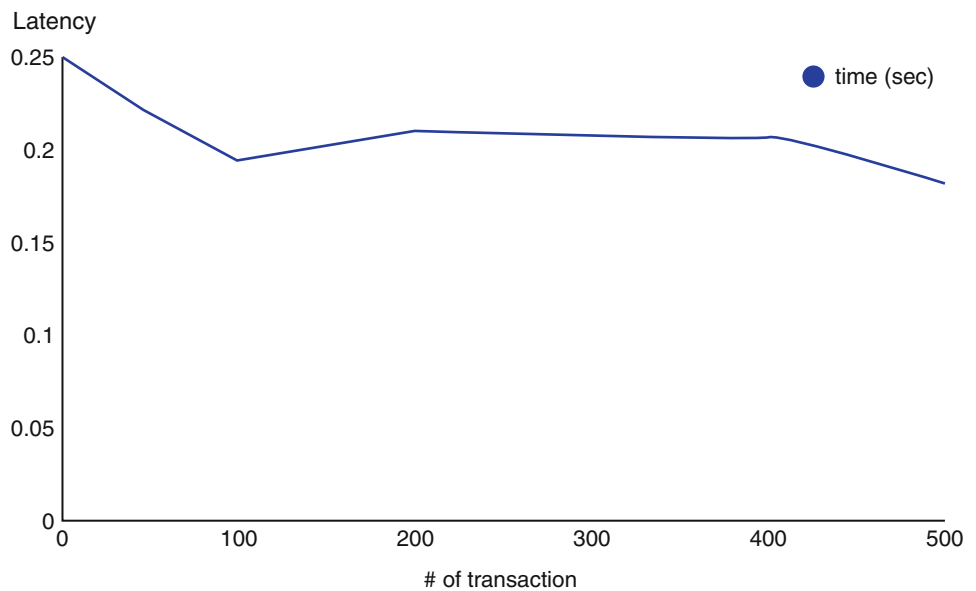
Number of transaction	Time (s)
1	0.27
100	19.5
200	41.07
300	61.98
400	82.72
500	90.5

with this network, which are specified in the smart contract. Then these readings are sent after checking the addresses and comparing the current data with the previous one for this meter, the decision is to establish a connection to execute this transaction depends on the result returned from the ethereum blockchain. At the start of the trials, the ethereum blockchain Private Network was configured. Then the proposed contracts are also published in that blockchain. Since this work is based on ethereum smart contracts, a new experiment is underway to evaluate the response time for reading the proposed smart contract. Table 56.1 shows the response time for writing smart contract data. The results showed that the speed of response time was relatively acceptable in the experiment. On the other hand, in a true application server environment, the hardware specifications will be much better than the device used in our experience. This will positively enhance the response time.

Figures 56.3 and 56.4 illustrates measures of response time and latency for meter writing readings into a blockchain network. These results were for a simple sample of readings from the meters and typed on the blockchain network.

56.6.1 DDoS Experiment Results

In our experiment, we studied the effect of the DoS attacks on the performance of servers connected with smart meters. The experiment was conducted by launching DoS attacks on servers, and then studied their strength against such attacks by analyzing their response time and their ability to communicate with the counting smart server while under attack. Packet generation tools can be used to build traffic or attack packets. For example, LOIC was used here. This tool performs a DoS attack by sending too many random data in the form of UDP, TCP or HTTP to the blockchain network to be dropped. The large numbers of packets are created per second and sent to servers. Figures 56.5 and 56.6 illustrate the effect of DDoS on transactions on a blockchain network, and the results of the experiments have also clearly shown that DDoS attacks had little effect on network performance. The responses from these servers to requests were very slow as

Fig. 56.3 Transaction response time**Fig. 56.4** Transaction latency

these requests did not affect the data itself but rather the speed of the blockchain network's response to the transactions received from the servers is decreased. Also, this delay in response time was not significant and had no effects on the blockchain network. The response times for each transaction were less than 0.25 ms before the DoS attacks. When testing DoS response times, they increased normally.

56.7 Conclusion

In this research we create a blockchain network that receives data from meters and sends it to servers, and then they send the data to the blockchain network. Besides, we implement every server with an address in the ethereum

network. It is through these addresses that data is sent and saved in the blockchain network. Blockchains are chains open to everyone who can read them. Transactions live in mempool before miner puts them in blocks. Generally, no one controls the blockchain so no one can go back to and change the data. As we saw in the results we cannot influence this network through the attacks, mainly because the networks are being paired and there is no need to trust the devices with each other, without a central point of failure. When miners are found, there is no need for central authority to tie one node with another or associate a user with access to another machine. In addition to studying the implementation of the blockchain network in smart meters and studying the effects of this proposal in terms of safety and security.

Fig. 56.5 Transaction response time with DDoS attack

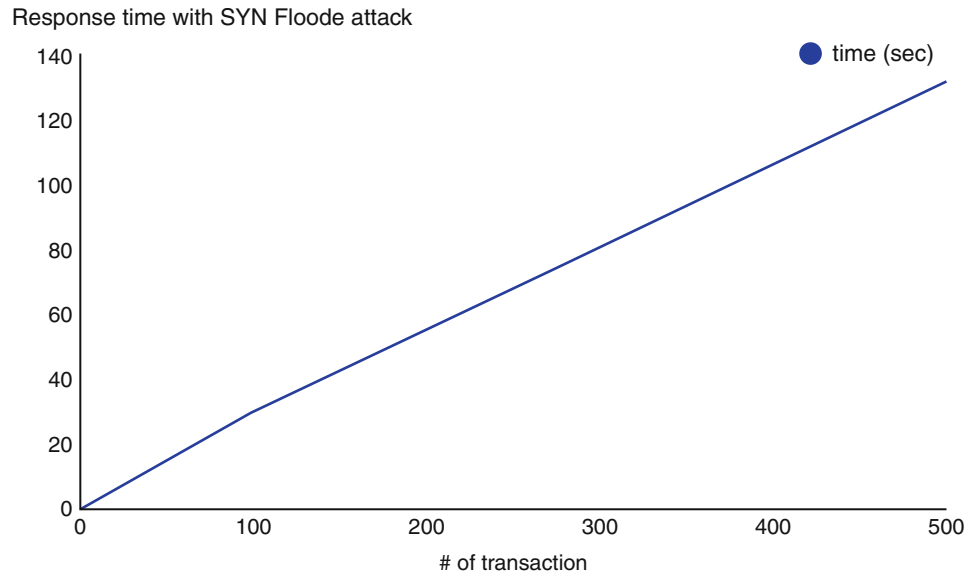
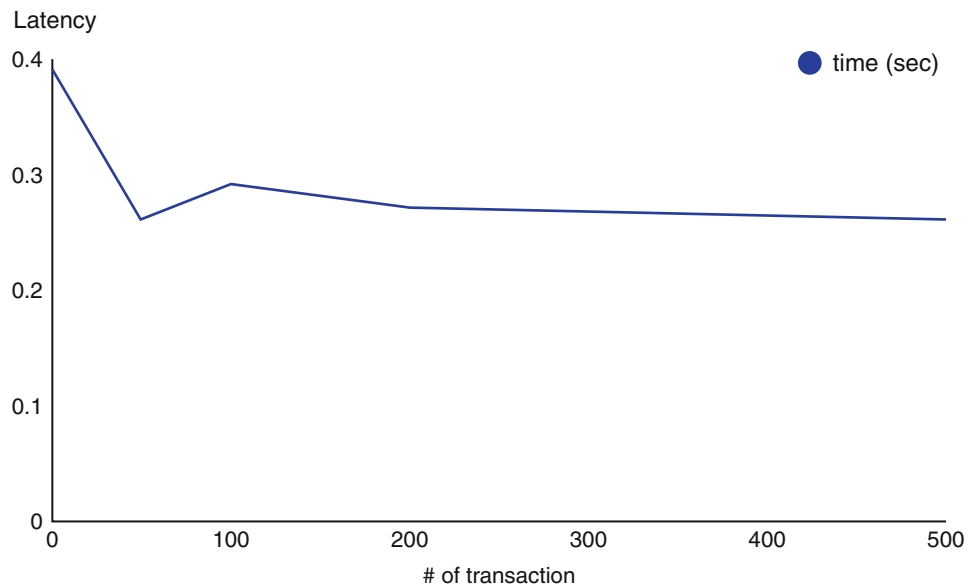


Fig. 56.6 Transaction latency with DDoS attack



References

1. P. Bramhe, A. Sarode, V. Bonde, R. Mankar, A. Kesharwani, S. Dhengale, Automatic electric meter reading using WiFi. *Int. Res. J. Eng. Technol* **6**(4), 343–346 (2019)
2. N. Chalaemwongwan, W. Kurutach, A practical national digital id framework on blockchain (NIDBC), in *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, (IEEE, 2018), pp. 497–500
3. A. Chore, P. Mali, D. Vyanjane, V. Karewar, IoT based smart electricity meter and billing system. *Int. Res. J. Eng. Technol.* **5**(4), 916–919 (2018)
4. X. Fan, C. Zhou, Y. Sun, D. Jinyang, Y. Zhao, Research on remote meter reading scheme and IoT smart energy meter based on NB-IoT technology. *J. Phys.* **1187**, 022064. IOP Publishing (2019)
5. R. Govindarajan, S. Meikandasivam, D. Vijayakumar, Cloud computing based smart energy monitoring system. *Int. J. Sci. Technol. Res.* **8**(10), 886–890 (2019)
6. W.-x. Lei, Y.-x. Jiang, W. Hong, A.-d. Xu, M. Zhe, W.-j. Hou, Y.-j. Yin, New features of automatic meter reading system: Based on edge computing, in *DEStech Transactions on Environment, Energy and Earth Sciences*, (ICEPE, 2019)
7. W. Mesbah, Securing smart electricity meters against customer attacks. *IEEE Trans. Smart Grid* **9**(1), 101–110 (2016)
8. O. Novo, Blockchain meets IoT: An architecture for scalable access management in IoT. *IEEE Internet Things J.* **5**(2), 1184–1195 (2018)
9. M. Pilkington, Blockchain technology: Principles and applications, in *Research Handbook on Digital Transformations*, (Edward Elgar Publishing, Cheltenham, 2016)
10. B. Sahani, T. Ravi, A. Tamboli, R. Pisal, IoT based smart energy meter. *Int. Res. J. Eng. Technol.* **4**(04), 96–102 (2017)
11. D. Vujičić, D. Jagodić, S. Randić, Blockchain technology, bitcoin, and ethereum: A brief overview, in *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, (IEEE, Piscataway, 2018), pp. 1–6

12. M. Warkentin, C. Orgeron, Using the security triad to assess blockchain technology in public sector applications. *Int. J. Inf. Manag.* **52**, 102090 (2020)
13. R. Zhang, R. Xue, L. Liu, Security and privacy on blockchain. *ACM Comput. Surv.* **52**(3), 1–34 (2019)
14. Z. Zheng, S. Xie, H. Dai, X. Chen, H. Wang, An overview of blockchain technology: Architecture, consensus, and future trends, in *2017 IEEE International Congress on Big Data (BigData Congress)*, (IEEE, 2017), pp. 557–564

Biometrics, Pattern Recognition and Classification

Using Machine Learning to Process Filters and Mimic Instant Camera Effect

57

Deirdre Chong, John Farhad Hanifzai, Hassan Adam, Jorge Garcia, and Jorge Ramón Fonseca Cacho

Abstract

In this paper we use Machine Learning to convert images taken with an iPhone camera and visually alter them to appear as if taken with a Leica Sofort Instant Camera, more commonly known as the Polaroid look. While such image filters already exist and are highly effective, they function using ad-hoc techniques. Our goal is to achieve similar results by having a model learn what the Polaroid look is on its own and how many image pairs are required to train it. We found that using linear regression we need, on average, 800 images before the model began displaying good consistent results while using Pix2Pix (Isola et al., Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134, 2017) (Conditional Adversarial Networks) and CycleGAN (Goodfellow et al., Generative adversarial nets. In Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems, vol 27, pp 2672–2680. Curran Associates, Inc., Red Hook, NY, 2014 [Online]. Available: <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>) (Generative Adversarial Networks) only required 500 images.

Keywords

Machine learning · Polaroid · Computer vision · Deep learning · Image processing

D. Chong · J. F. Hanifzai · H. Adam · J. Garcia
 J. R. Fonseca Cacho (✉)
 Department of Computer Science, University of Nevada, Las Vegas,
 Las Vegas, NV, USA
 e-mail: Jorge.FonsecaCacho@unlv.edu

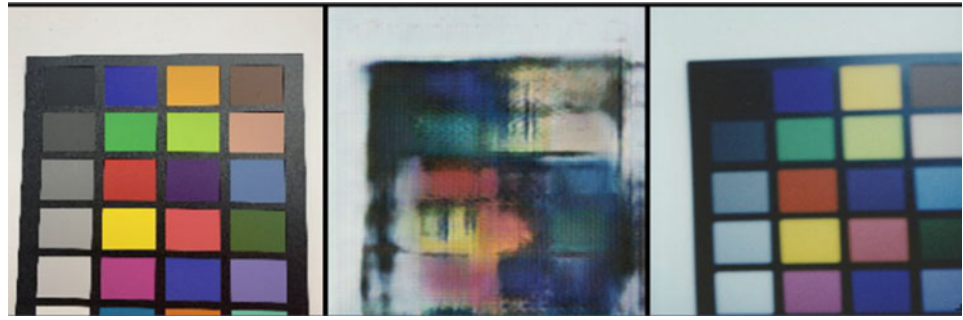
57.1 Introduction

Machine Learning has seen a massive demand in computer applications over the past few years. From self-driving cars to predicting the future, Machine Learning makes it possible for computers to solve problems more quickly and efficiently than humans can [1]. Our research aims to apply Machine Learning in image processing. The potential of Machine Learning in image processing is limitless [2]. Machine Learning can be implemented in image processing in various ways, including photo enhancement [3]. Image-to-image translation has been implemented in intriguing ways, such as converting a stock photo to a canvas, mimicking Van Gogh and Monet [4]. This research aims to apply Machine Learning to develop a model that will convert a standard unprocessed photo into a photo that visually looks as if taken with an instant camera, Leica Sofort. This camera has a unique look that many people know as the ‘Polaroid’ look (Fig. 57.3). There are pre-existing image filters that replicate instant camera images, but none accurately capture all the differences and subtleties [5]. Image filters made with Machine Learning can replace outdated image processing approaches and lead endless image processing possibilities with the right training data set [6].

57.2 Background

Usually, image-to-image translation utilize pixel-to-pixel models to predict the corresponding output image. Translation in this sense is transposing from a representation or a format to another, this could either be an RGB image, edge maps, label maps, etc. These image translations can be implemented with many different types of algorithms, one is linear regression. This research uses linear regression models

Fig. 57.1 Pix2Pix Real
A—Generated B—Real B [4]



to parse each of the RGB values per pixel of the paired image data sets, as per the following,

$$\begin{aligned} R_{ij} &= w_1oldR_{ij} + w_2oldG_{ij} + w_3oldB_{ij} + w_0 \\ G_{ij} &= w_1oldR_{ij} + w_2oldG_{ij} + w_3oldB_{ij} + w_0 \\ B_{ij} &= w_1oldR_{ij} + w_2oldG_{ij} + w_3oldB_{ij} + w_0 \end{aligned}$$

Given a pixel at coordinate (i,j), we independently compute three regressions for each color. Each color has three weights that represent the old RGB value of the original pixel and the bias. The three weights of each of model are independent of the weights of the other models. The reason RGB values are utilized to compute each color separately is to try to find a relation between the other colors and how each individual color is affected.

Previous image-to-image translation research has shown that it is possible to predict pixel structure using convolutional neural networks (CNNs) [7]. These solutions work by having an image with a missing region, processing it through the convolutional neural network and regress the missing pixel values. These solutions are trained in a completely unsupervised manner. A similar image-to-image translation solution can be seen using conditional adversarial networks as seen on the paper. Conditional adversarial networks efficiently learn the mapping and learn a loss function to train the model [4].

Pix2Pix aims to solve the same problem of translating an input image into its corresponding output. We explore the use of generative adversarial networks (GANs) for these tasks. GANs are suited in the conditional setting because they can automatically learn a loss function to classify if an image is real and at the same time, train a generative model to minimize this loss. Figure 57.1 shows the early stages of training the Pix2Pix generative adversarial network.

It is worth noting that the results that will eventually come from processing images through our several models, have to be reviewed by a human to interpret and compare results. Even though this is inherently objective, as there is no algorithmic way to tell if the image transformations are effective, we established a criteria to determine if an image passed the transformation or is rejected. For this criteria, we consider attributes like:

- Image maintains its colors after transformations
- Image is still discernible without major distortions
- Image still resembles original without loss of detail
- Image still maintains most of the original colors
- Image kept similar proportions

57.3 Methods

Machine Learning requires enough data to teach the computer how to convert the given input into the desired output. The dataset is split into two parts: one that contains normal images and ones that contain our desired output. The more examples we have, the better it can detect subtle unexplainable patterns.

For this research we devised a very specific workflow. The first step was getting a dataset of images that would eventually go through the regression model as a first attempt at image transformation. The regression model would then take the R, G, or B values from the whole dataset. We would then need a new script to extract the RGB values from these images in the dataset, and lastly, the linear regression script.

As mentioned before, the main step for creating a dataset was downloading 6000 images. We requested downloads from the Bing search engine's API with the help of libraries like OpenCV, to test if the images were valid or corrupted, otherwise they would have to be deleted as a first step into normalization. Having created a big enough batch of images to process, we then had a new script to resize all images to 300×300 pixels, and we edited the images by introducing contrast and color changes and altered them into different styles using image processing tools like ImageMagik and OpenCV.

```
#Dataset config snippet

# train and test split for dataset
if argv.split is not None and
    argv.split < 1:

    split_ratio = argv.split
else:
    split_ratio = .75

img_count = 0
```



```

for photo in img_A:
    if photo.lower().endswith(
        ('.png', '.jpg', '.jpeg')):

        img_count += 1

img_split = math.ceil(img_count
    * (split_ratio))

...

exit()

```

This step was necessary to reduce film cost by estimating the minimum number of images required for the Machine Learning algorithm to produce useful outputs. After we found the optimal number of images, we collected Polaroid images and digital images, then matched the pixels using Adobe Photoshop.

Afterwards, we read the RGB values of each photo and save them at a main NumPy array with each image image array inside it.

#RGB Numpy Array creation code snippet

```

image = np.array([])
r_ds = np.array([])
g_ds = np.array([])
b_ds = np.array([])

currentDirectory = os.getcwd()

for filename in os.listdir("./"):
    if filename.endswith(".jpg"):
        directory = ''
        fileDir = cv2.imread(os.path.join
            (directory, filename))
        #print('Working on file:', fileDir)
        img = cv2.cvtColor
            (fileDir, cv2.COLOR_BGR2RGB)/255

        r_val = img[:, :, 0]
        g_val = img[:, :, 1]
        b_val = img[:, :, 2]

        r_ds = np.append(r_ds, r_val)
        g_ds = np.append(g_ds, g_val)
        b_ds = np.append(b_ds, b_val)

        index += 1
        if index == 1000:
            break;
        continue

np.save('r.npy', r_ds)
np.save('g.npy', g_ds)
np.save('b.npy', b_ds)

```

Finally, we used our polaroid dataset on machine learning models like Linear Regression, GANs (Generative Adversarial Networks) such as Cycle-GANs, and Pix2Pix GANs. Generative Adversarial Networks, more broadly the adversarial nets frameworks were chosen because they generative models sidestep the difficulties they usually convey, like the difficulty to approximate untraceable probabilistic computations that come from estimations. This framework is interesting because it places a model against an adversary model that learns to determine whether a sample is from the first model or from dataset. This competition between the two models usually produces improvements in their results [8].

57.4 Test Results

As our initial test, we applied linear regression to predict images on our test datasets. The paired images are resized to the same resolution and RGB values are extracted from the image as mentioned previously. These values are then fitted through our linear regression script. Figure 57.2 shows the sample images and the result using linear regression for 800 training images.

Linear regression R G B

```

print('loading ....')
x1 = np.load('./numpy/r.npy')
x2 = np.load('./numpy/g.npy')
x3 = np.load('./numpy/b.npy')

y1 = np.load('./numpy/r1.npy')
y2 = np.load('./numpy/g1.npy')
y3 = np.load('./numpy/b1.npy')

print('dataframe ....')
data = pd.DataFrame
    (np.c_[x1,x2,x3,y1,y2,y3])

print("Data: \n",data)
#print('reshaping ....')

X = data[[0,1,2]]
print("X: \n", X)

y = data[[3]]
print("y: \n", y)

print('splitting ....')
X_train, X_test,y_train,y_test = t
    rain_test_split(X,y,test_size=.
        3,random_state=1)

print('fitting ....')
simplemodel = LinearRegression()
simplemodel.fit(X_train, y_train)

```

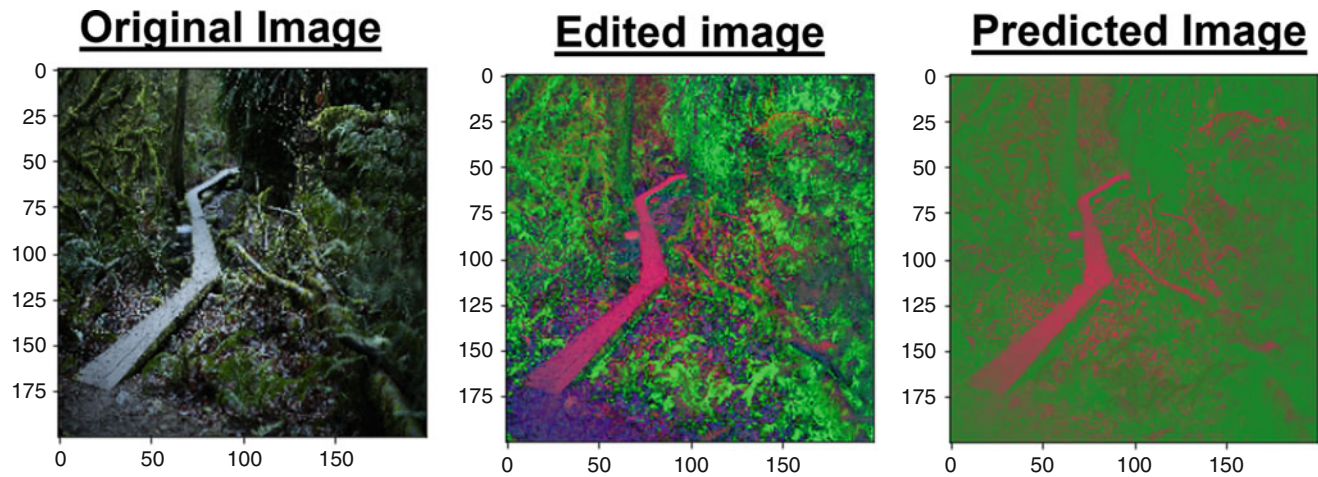


Fig. 57.2 Side-by-side of a sample image that was then modified and fed to our Linear Regression model and the resulting image

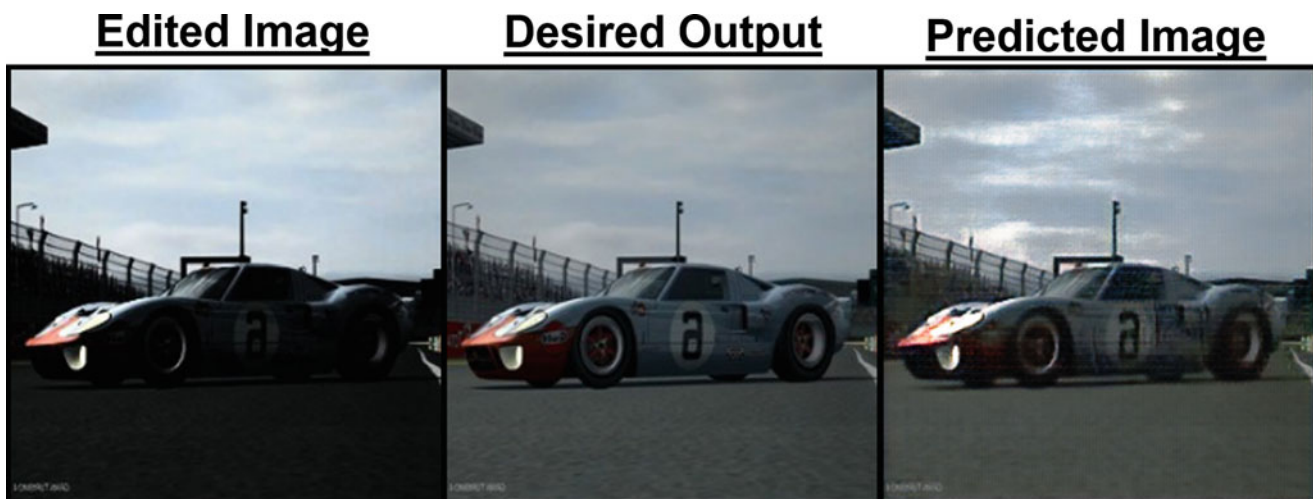


Fig. 57.3 Pix2Pix on contrast test dataset [4]

The next step on our test is to use Pix2Pix GAN on our test dataset. The contrast is edited with ImageMagik and used as the original input image. The model does a good job recovering back details that were lost when the contrast was heightened. Figure 57.3 shows the results of the model after 50 epochs trained with 500 paired images.

57.5 Results

The trial run with digitally edited images resulted in successful prediction in Linear Regression with 800 images. Machine Learning models like Pix2Pix, predicted the edited styles with 500 images.

Since Polaroid film was scarce and expensive, we used 150 Instant Camera Images with a corresponding digital image. The images were then fitted through the Linear Regression Model.

Figure 57.4 shows the prediction using linear regression using digital image and polaroid image datasets. The results alter the color to appear duller which slightly mimics the polaroid image as shown in Fig. 57.5.

Then we used GANs on the same dataset and compared the results. We used the Pix2Pix GAN and CycleGAN on the 150 Instant Camera Images. Figure 57.6 shows the image prediction using Pix2Pix GAN. Figure 57.7 shows the image prediction using CycleGAN.

57.6 Findings

Our findings predict accurate results with 500 images on our edited datasets. Our datasets with 150 instant camera images showed promising results with Linear Regression and Generative Adversarial Networks. GANs work better in predicting and finding the right mapping. Linear Regres-

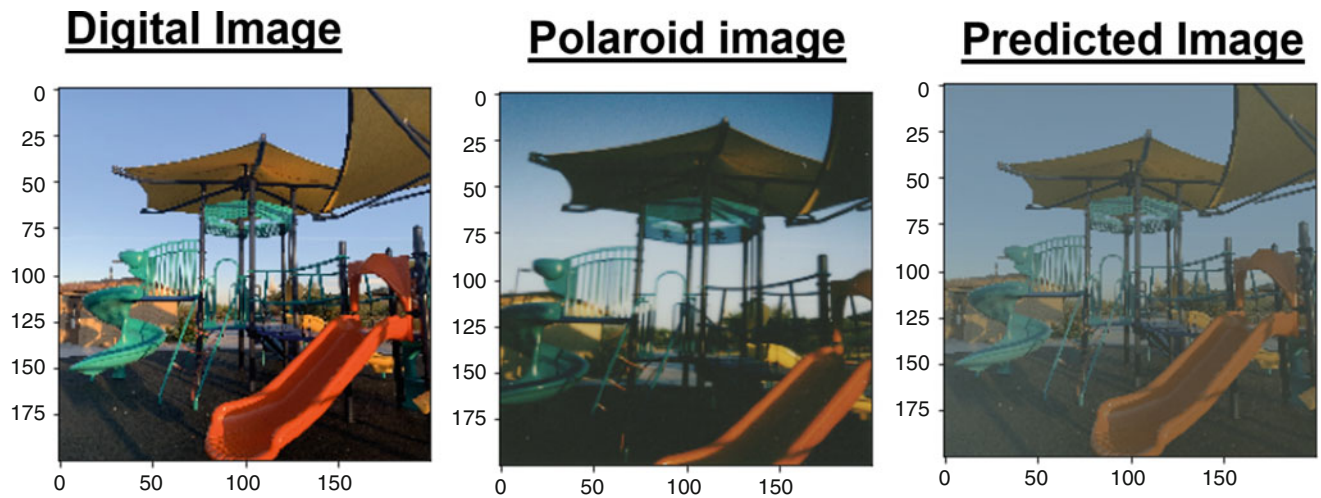


Fig. 57.4 Linear Regression side-by-side on our dataset

Fig. 57.5 Color comparison between original, polaroid, and the predicted data set

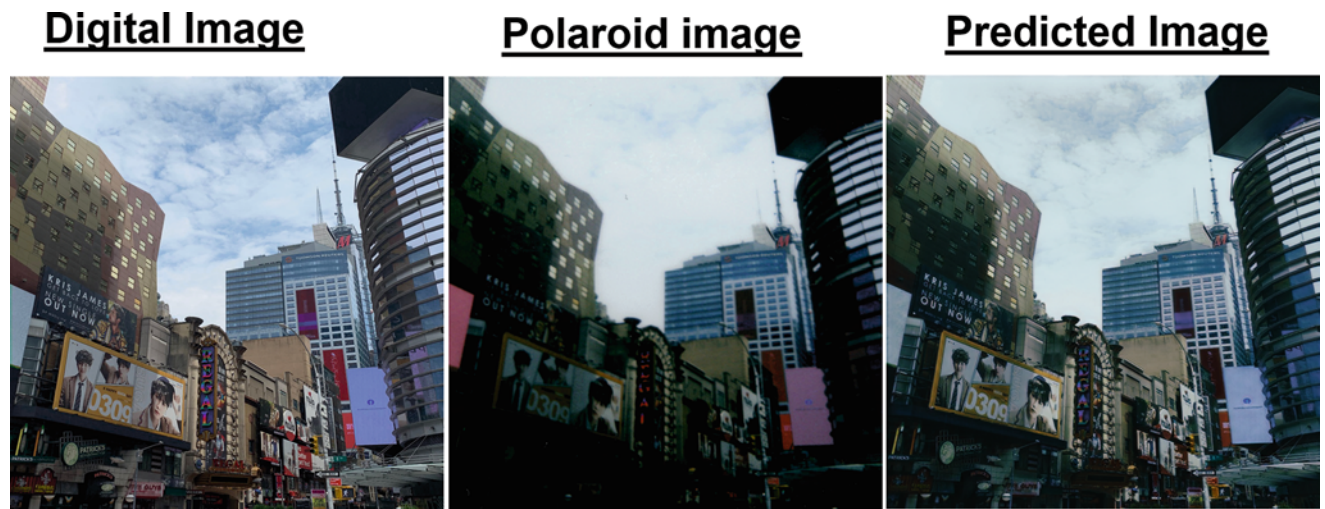


Fig. 57.6 Running Pix2Pix [4] on our dataset

sion is very limited and GANs can pick up more efficient mappings by training a loss function. Our research found that GANs predicts colors with much higher accuracy and at least 500 images are necessary to create accurate results. It is worth noting that an accurate result might be subjective, but we interpret it as one that any untrained person would not be able to distinguish by simply looking at a sample.

57.7 Conclusion

Machine Learning has a lot of potential applications in image processing. Our research found accurate results with GANs like Pix2Pix and CycleGAN. Our research showed that Machine Learning can be implemented to create better image processing solutions. Machine Learning can find new ways to

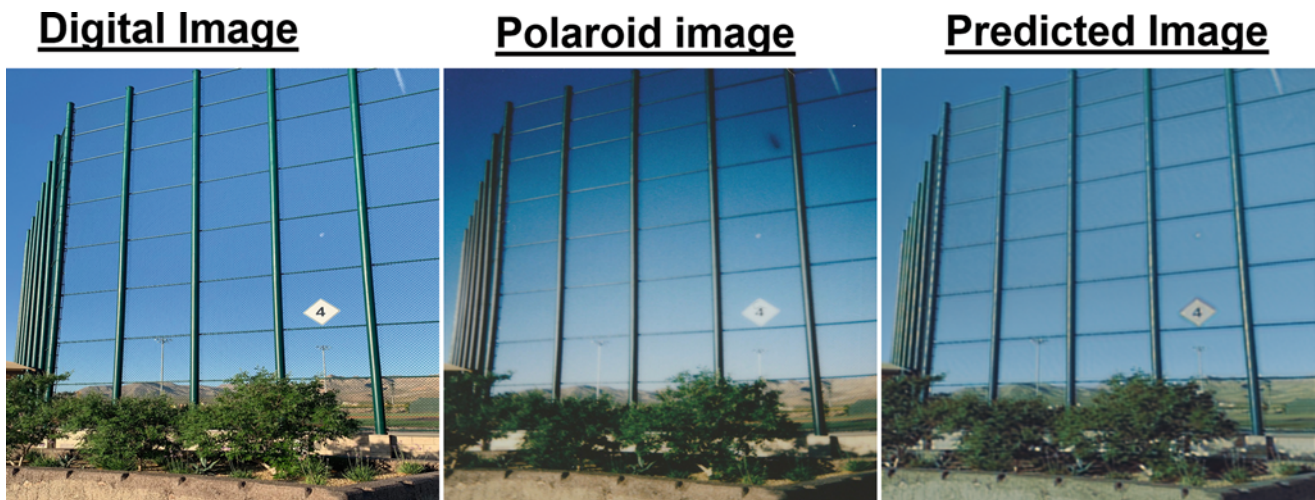


Fig. 57.7 Running CycleGAN [3] on our dataset

mimic subtle differences and create a new approach to image processing [6].

The size of our datasets limited the results of our research. Our research dataset only contained 150 images, due to the difficulty of taking pictures using a Polaroid and scanning them in a high amount. By using more images, we anticipate an increase in the accuracy of the machine learning models. Other machine learning models also have yet to be tested to conclude the best model for this effect. There is a lot of potential using machine learning to create new filters and image effects. We can experiment with our datasets and create colder effects. Other aspects of photos can also increase accuracy, such as brightness, contrast, grain, etc.

Even though our experiment was limited, we can see a promising result. The result of our Linear Regression model showed cold alterations of the colors used while GANs replicated the Polaroid effect very well. In future work, we intend to attempt ensemble learning to combine results of these and other Machine Learning models to improve on these results, along with increasing the size of our dataset. Because reproducible research [9] is important to allow others to review and improve on our results, we are making our source code and data publicly available at UNLV and Github repositories.

Acknowledgments We would like to thank UNLV's Center for Academic Enrichment and Outreach and the Office of Undergraduate Research for funding this research. This material is based upon work supported by the National Science Foundation under Grant No. 1625677. This material was supported by the U.S. Department of Education: Asian American and Native American Pacific Islander-Serving Institutions (AANAPISI) program, Grant No. P031150019.

References

1. D. Howard, D. Dai, Public perceptions of self-driving cars: the case of Berkeley, California, in *Transportation Research Board 93rd Annual Meeting*, vol. 14(4502), 1–16 (2014)
2. J. Leban, Image recognition with machine learning on python, image processing. <https://towardsdatascience.com/image-recognition-with-machine-learning-on-python-image-processing-3abe6b158e9a>. Accessed 20 October 2020
3. J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2223–2232
4. P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1125–1134
5. Easybluecode, Instalab iOS app. <https://apps.apple.com/us/app/instalab-instant-films/id1113395996>. Accessed 20 October 2020
6. L.J. Spreeuwens, Image filtering with neural networks: applications and performance evaluation (1994)
7. D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: feature learning by inpainting, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems*, vol. 27, ed. by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Curran Associates, Inc., Red Hook, NY, 2014), pp. 2672–2680 [Online]. Available: <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>
9. J.R.F. Cacho, K. Taghva, The state of reproducible research in computer science, in *17th International Conference on Information Technology–New Generations (ITNG 2020)* (Springer, New York, 2020), pp. 519–524

Benchmarking Accuracy and Precision of the Convolutional Neural Networks for Face Recognition on Makeup and Occluded Images

Stanislav Selitskiy, Nikolaos Christou, and Natalya Selitskaya

Abstract

We benchmark established state-of-the-art Convolutional Neural Network (CNN) models for face recognition in the real-life conditions on the novel data set with makeup and occlusions. Strength and weaknesses of different CNN implementation architectures and degree of depth on particular types of makeup and occlusions are identified in the context of the selecting complementary models for the ensemble use. A practical approach of isolating uncertainty of the model verdicts trustworthiness in order to boost precision is investigated.

Keywords

Deep learning · Biometrics · Face recognition · Makeup · Occlusion · Disguise · Spoofing · Source code · Uncertainty · Reliability

58.1 Introduction

In recent years Machine Learning (ML) and Artificial Neural Networks (ANNs) technologies have been successfully developed for pattern recognition in the presence of noise and natural variations. The efficient ML solutions providing estimates of uncertainty in pattern recognition have been developed for trauma healthcare [1–3], and collision avoid-

ance [4]. In particular, Deep Learning (DL) solutions developed for face recognition have achieved high accuracy on benchmark problems [5, 6]. However, there are still edge-case conditions that can significantly reduce the recognition accuracy of the existing ML solutions, demonstrating high accuracy on benchmark and real data [7–9]. Such cases require experimental exploration of the limitations of the existing ML technologies, as shown, e.g. in [10–12].

In this paper, we aim to study the influence of disruptive and spoofing activities which significantly affect face recognition. The study is undertaken on a new benchmark set of facial images, including visual obstacles such as makeup, face masks, headdress, and eyeglasses.

The influence of makeup and occlusion on face recognition has been studied in early work discussed in [13–15] and later in [16–20]. The related benchmark data sets have been collected for the face recognition study in [21–23].

The majority of the makeup and occlusion containing benchmark data sets are composed of the internet scraped images, which allows collecting data for a large number of subjects. However, they have a quite narrow, sparse and inconsistent variety of makeup types, facial expressions, face orientations, image qualities and dimensions per person [24–26].

The motivation for creating the novel publicly available BookClub data set was an attempt to fill in the apparent gap in the makeup and occlusions data sets with the one containing high quality and resolution images, featuring sophisticated original artistic makeup and non-trivial occlusions.

As most of the cited work concentrated on the effects of makeup and occlusions on face recognition using engineered features algorithms, the behaviour of the modern CNN architectures was investigated.

The paper is organised as follows. Section 58.2 presents a detailed description of the BookClub data set. Section 58.3 lists how the data set was structured in the experiments and how the models were applied; Sect. 58.4 presents the

S. Selitskiy (✉) · N. Christou
University of Bedfordshire, School of Computer Science, Luton, UK
e-mail: vitaly.schetinin@study.beds.ac.uk;
nikolaos.christou@study.beds.ac.uk

N. Selitskaya
Kennesaw State University, College of Science and Mathematics,
Kennesaw, GA, USA
e-mail: nselitskaya@kennesaw.edu

obtained results, and Sect. 58.5 draws practical conclusions from the results and states directions of the research of not yet answered questions.

58.2 Data Set

The novel BookClub artistic makeup data set contains images of 21 subjects. Each subject's data may contain a photo-session series of photos with no-makeup, various makeup (MK session suffix), and images with other obstacles for facial recognition, such as wigs (HD), glasses (GL), jewellery, face masks (FM), or various types of headdress (HD). Overall, the data set features 37 photo-sessions without makeup or occlusions, 40 makeup sessions, and 17 sessions with occlusions. Each photo-session contains circa 168 RAW images of up to 4288×2848 dimensions (available by request) of six basic emotional expressions (sadness, happiness, surprise, fear, anger, disgust), a neutral expression, and the closed eyes photo-shoots taken with seven head rotations at three exposure times on the off-white background.

The photos were taken on Nikon D5000 camera with Sigma 50 mm lens with no flash in the studio with soft lighting on the white cloth background. Majority of the sessions were taken with 200 ISO speed rating, auto-focus, $f/3.2$ aperture priority, and three exposures with 1.7 stop difference, which resulted in circa 1/15, 1/50, 1/160 seconds exposures. The master photosest storage format is NEF, a 16-bit colour space format up-scaled from the 12-bit colour depth sensor data.

Default publicly available downloadable format is a JPEG of the 1072×712 resolution. The subjects' age varies from their twenties to sixties. Race of the subjects is predominately Caucasian and some Asian. Gender is approximately evenly split between sessions [27].

The photos were taken over the course of 2 months. A few sessions were done later, and some subjects posed at multiple sessions over several week intervals in various clothing with changed hairstyles.

58.3 Experiments

In [28], on the AlexNet CNN model example, it was shown that even state-of-the-art machine learning algorithms are prone to face recognition errors on particular types of makeups, occlusions, and spoofed personalities. For the follow-up enquiry, other state-of-the-art CNN models of various degree of deepness and connection structure, such as VGG19, GoogLeNet, Resnet50, Inception v.3, InceptionResnet v.2, were experimented with.

To emulate the real-life conditions, when makeup or other occlusions of the subjects may not be predicted and included

into training sets beforehand, non-makeup and non-occlusion sessions were selected into the training set amounted to roughly 6200 images. The high-level recognition accuracy was calculated as a ratio between the number of correctly identified images and the whole number of images in the session (similarly, the "guess score" for misidentified sessions).

The AlexNet model consists of 25 layers and takes as input images scaled to 277×277 dimensions. The VGG19 model has 47 layers, GoogLeNet – 144, Resnet50 – 177, and they take 224×224 scaled images as input. The Inception v.3 contains 315 layers, and InceptionResnet v.2 – 824, both taking 299×299 scaled images as input. All of the experimented with models, but AlexNet and VGG19 models, have Directed Acyclic Graph (DAG) architecture. "Adam" learning algorithm with 0.001 learning coefficient, mini-batch size 64 parameters are used for training. Depending on the rate convergence of the models, 10, 20, 30 epochs were used.

For the precision improvement experiments, for each subject with more than one non-makeup session, one session was set aside for the post-training assessment of the wrong guess soft-max activation distribution, which amounted to 1653 images. The hypothesis was tested on makeup and occluded images, which amounted to 9492 number, that the prior wrong guess activation distribution can be used to estimate trusted accuracy and precision at the given confidence level.

In such a way, the uncertainty of the low-score verdicts is isolated in the "unidentified" class, decreasing the trusted accuracy and boosting the precision of the trusted verdicts.

58.4 Results

Introducing more complex and deeper models into experiments has helped increase the hypothetical ensemble's accuracy, would they be used together. However, a sizable number of sessions remained misidentified either by all or majority of the tested models, see Table 58.1a, b and Fig. 58.1.

Another set of the "problematic" sessions, see Table 58.2a, b, exposes "blind spots" of particular CNN architectures. Some models demonstrate very high, almost ideal accuracy for such sessions, while others have negligibly low or virtually zero accuracies.

Particularly, VGG19 model had problems with the artificial white wig that other models easily recognised. Inception v.3, which was the most accurate and reliable model overall, had few "blinders" on simple makeups, easily recognisable by even simpler models. Resnet50 failed to recognise painted realistic human faces. Furthermore, GoogLeNet failed on face masks which other DAG models solved recognition. Inception v.3 and InceptionResnet v.2 solved recognition of the

Fig. 58.1 Image class examples misidentified by all CNN models, left to right and top to bottom: Subj.7, Sess.MK2; Subj.10, Sess.MK4; Subj.14, Sess.HD1, Subj.21, Sess.MK1



Table 58.1 Misidentified image classes by all or majority of the CNN models. Session accuracy by the models

(a)			
Session	AlexNet	VGG19	GoogLeNet
S1HD1	0.1325	0.6747	0.5783
S1MK3	0.0000	0.6989	0.1477
S2HD1	0.3742	0.0000	0.0204
S4GL1	0.0201	0.0000	0.1042
S4MK1	0.0000	0.1043	0.0123
S5MK3	0.0484	0.0364	0.2424
S7MK2	0.3046	0.0000	0.0000
S7FM1	0.0000	0.0000	0.0000
S10MK2	0.0000	0.1018	0.1437
S10MK3	0.0000	0.0000	0.0000
S10MK4	0.1258	0.0000	0.0000
S14HD1	0.0368	0.0000	0.1963
S20MK1	0.0000	0.7048	0.0000
S21MK1	0.0000	0.0000	0.0000
(b)			
Session	Resnet50	Inception3	InceptRes2
S1HD1	0.0000	0.0000	0.0000
S1MK3	0.0000	0.0000	0.0739
S2HD1	0.9932	0.2109	0.0000
S4GL1	0.0000	0.8472	0.0069
S4MK1	0.0368	1.0000	0.0000
S5MK3	0.1455	0.1515	0.9939
S7MK2	0.0000	0.0402	0.1149
S7FM1	0.0000	0.4568	0.7531
S10MK2	0.0298	0.0120	0.7066
S10MK3	0.0178	0.0000	0.0000
S10MK4	0.0000	0.0000	0.0599
S14HD1	0.2025	0.3436	0.0000
S20MK1	0.0000	0.0060	0.0060
S21MK1	0.0000	0.0000	0.0000

light theatrical type makeup that was problematic for other models. VGG19, while failing on many easy cases, uniquely recognised heavily painted over faces with contrast pigments. GoogLeNet and Resnet50 were particularly successful with recognition in the presence of wigs and dark glasses, and Inception v.3 – for face mask recognition.

In the presence of mutually contradicting verdicts with soft-max activation scores in the full 0.0–1.0 range, a simple, fast, and practical method of the A/B test was investigated to obtain the trusted verdicts using prior, training time data. A distribution shape similarity (see Fig. 58.2) can be noted for the training-time soft-max activations, which drove the classification verdicts, for wrongly classified non-makeup images and the test-time distributions for images with makeup and occlusions.

The rigorous inferential statistics hypotheses testing algorithms for the goodness of fit and variance homogeneity, such as Kolmogorov-Smirnov or Mann-Whitney-Wilcoxon, do not show a high probability of these samples belonging to the same population. However, we verify another hypothesis that the trusted accuracy and precision, based on A/B thresholds calculated for given confidence levels for the posterior, obtained from the test-time distributions serve as a lower bound for the trusted accuracy and precision calculated for the prior, training-time distributions. I.e. the prior trusted accuracy and precision are no worse than the posterior. Which can be seen is either satisfied for a number of the tested models and confidence levels or close for practical purposes, see Tables 58.3a, b and 58.4a, b.

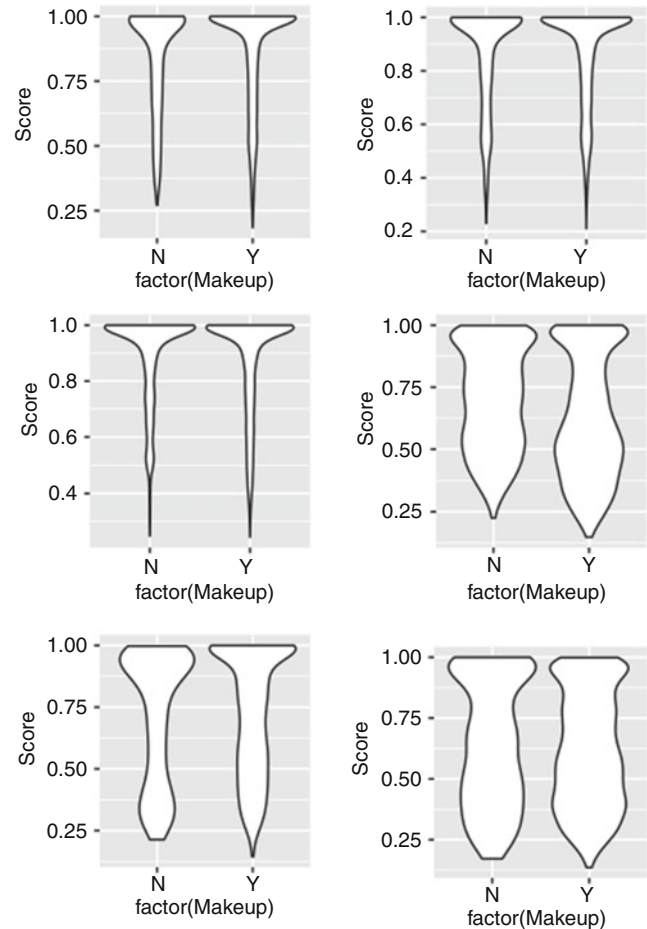
Unconstrained accuracy of the model is calculated as a ratio of the number of correctly identified test images to a number of all test images:

$$Accuracy = \frac{N_{correct}}{N_{all}}$$

Table 58.2 Image classes identified by some of the CNN models with high accuracy, and some – with low. Session accuracy by the models

(a)			
Session	AlexNet	VGG19	GoogLeNet
S1HD2	0.9939	0.0061	1.0000
S1MK1	0.9583	0.9941	1.0000
S1MK7	0.7371	0.9600	0.5429
S1MK8	0.2470	0.1988	0.9880
S2GL1	0.3151	0.0479	0.9658
S5MK1	0.0000	0.0538	0.0000
S5MK2	0.0000	0.0114	0.0000
S6MK1	0.0000	0.2036	0.0059
S7MK1	0.0000	0.0000	0.5950
S7FM2	0.0000	0.1065	0.0000
S7FM3	0.1428	0.0179	0.0000
S10MK1	0.9702	0.0000	0.6429
S12MK1	0.8935	0.4464	0.0417
S12MK2	0.8494	0.0000	0.6084
S14GL1	0.7725	1.0000	0.9940
S14MK1	0.8795	0.8253	0.7349
S15MK1	0.9091	0.0000	0.0061
S17MK1	0.3455	0.0484	0.8849
(b)			
Session	Resnet50	Inception3	InceptRes2
S1HD2	0.8598	1.0000	0.0000
S1MK1	0.8869	0.0059	0.0000
S1MK7	0.0343	0.0000	0.0000
S1MK8	0.8675	0.0000	0.0000
S2GL1	1.0000	1.0000	0.0000
S5MK1	0.4012	1.0000	0.9761
S5MK2	0.0178	0.8639	1.0000
S6MK1	0.0000	0.8743	0.6407
S7MK1	0.0000	0.7857	0.9762
S7FM2	0.0000	0.8521	0.8994
S7FM3	0.1191	0.9048	1.0000
S10MK1	0.9583	1.0000	1.0000
S12MK1	1.0000	0.8810	0.0000
S12MK2	0.0000	0.4699	0.0000
S14GL1	0.9940	1.0000	0.0000
S14MK1	0.8795	1.0000	0.0000
S15MK1	0.0101	1.0000	0.1515
S17MK1	0.7091	1.0000	1.0000

The trusted accuracy at the given confidence level is calculated as a ratio of the number of the correctly identified test images with soft-max activation (viewed as a probability P of the image belonging to the class) greater than the desired confidence percentile and number of incorrectly identified test images with soft-max activation smaller than the desired confidence percentile, relative to a number of all test images:

**Fig. 58.2** Violin plots of the prior (selected non-makeup sessions) and posterior (sessions with makeup and occlusions) distributions of the soft-max activation score that drove the classification verdict for the wrongly identified images for AlexNet, VGG19, GoogLeNet, Resnet50, Inception v.3, InceptionResnet v.2 models (left to right and top to bottom)

$$PCTL \text{ Accuracy} = \frac{N_{correct:P>PCTL} + N_{wrong:P \leq PCTL}}{N_{all}}$$

The trusted precision is calculated similarly to the trusted accuracy as the ratio of the correctly identified test images with soft-max activation greater than the desired confidence percentile to the number of all identified as particular class test images with soft-max activation greater than the desired confidence percentile:

$$PCTL \text{ Precision} = \frac{N_{correct:P>PCTL}}{N_{correct:P>PCTL} + N_{wrong:P>PCTL}}$$

The trusted recall is calculated as the ratio of the correctly identified test images with soft-max activation greater than the desired confidence percentile to the number of all correctly identified test images:

$$PCTL \text{ Recall} = \frac{N_{correct:P>PCTL}}{N_{correct}}$$

Table 58.3 Unconstrained accuracy, trusted accuracy and precision of the test images calculated based on the confidence level thresholds of the prior distributions of the wrongly identified images for selected non-makeup sessions

(a)			
Metric	AlexNet	VGG19	GoogLeNet
Accuracy	0.3643	0.3833	0.4827
75% Pr CL threshold	0.9972	0.9995	0.9995
90% Pr CL threshold	0.9999	0.9999	0.9999
75% Pr CL accuracy	0.6638	0.6769	0.8217
90% Pr CL accuracy	0.6989	0.6643	0.8009
75% Pr CL precision	0.5507	0.6404	0.8423
90% Pr CL precision	0.6454	0.7991	0.9526
75% Pr CL recall	0.6639	0.4750	0.7870
90% Pr CL recall	0.4746	0.2372	0.6288
(b)			
Metric	Resnet50	Inception3	InceptRes2
Accuracy	0.5457	0.5966	0.5529
75% Pr CL threshold	0.8954	0.9455	0.9995
90% Pr CL threshold	0.9741	0.9788	0.9999
75% Pr CL accuracy	0.7700	0.7581	0.7786
90% Pr CL accuracy	0.7570	0.7263	0.7644
75% Pr CL precision	0.8365	0.8150	0.8765
90% Pr CL precision	0.8953	0.8285	0.9665
75% Pr CL recall	0.7679	0.7982	0.7383
90% Pr CL recall	0.6751	0.7146	0.6329

Table 58.4 Verification trusted accuracy and precision of the test images calculated based on the confidence level thresholds of the posterior distributions of the wrongly identified images for sessions with makeup and occlusions

(a)			
Metric	AlexNet	VGG19	GoogLeNet
75% Pos CL threshold	0.9995	0.9980	0.9978
90% Pos CL threshold	0.9999	0.9999	0.9998
75% Pos CL accuracy	0.6866	0.6649	0.7910
90% Pos CL accuracy	0.6943	0.6817	0.8189
75% Pos CL precision	0.5920	0.5989	0.7653
90% Pos CL precision	0.7078	0.7159	0.8780
75% Pos CL recall	0.5844	0.5416	0.8328
90% Pos CL recall	0.3437	0.3650	0.7361
(b)			
Metric	Resnet50	Inception3	InceptRes2
75% Pos CL threshold	0.8654	0.9805	0.9978
90% Pos CL threshold	0.9824	0.9997	0.9998
75% Pos CL accuracy	0.7723	0.7249	0.7880
90% Pos CL accuracy	0.7486	0.6288	0.7743
75% Pos CL precision	0.8263	0.8297	0.8331
90% Pos CL precision	0.9076	0.8897	0.9139
75% Pos CL recall	0.7871	0.7103	0.8128
90% Pos CL recall	0.6483	0.4707	0.6926

58.5 Discussion and Future Work

Testing the diverse set of state-of-the-art CNN models of the different complexity levels, still demonstrate their shortcomings in the real-life style training and testing for images with makeup and occlusions. There is a subset of the problematic types of photo-sessions in the benchmark data set on which all tested models failed. Especially fascinated are those cases when some photo-sessions generate opposite verdicts for different models. In such cases, it is seen that particular models solved difficult problems, such as dark glasses or wigs, while failing on other tasks that are solved by other simpler models. Future work in finding explanations which parts of particular architectures are responsible for weak and robust behaviour would be beneficial in this context.

The presented precision increasing technique that uses prior, training-time data cannot guarantee strict confidence level compliance. However, it can still be a viable way of increasing the robustness and trustworthiness of the applications sensitive to false-positive errors. A meta-learning supervisor neural network approach may be more prospective and accurate to learn a more complex and multidimensional dependency of the model robustness from the softmax activation verdicts.

References

1. V. Schetinina, L. Jakaite, W.J. Krzanowski, Prediction of survival probabilities with Bayesian decision trees. *Expert Syst. Appl.* **40**(14), 5466–5476 (2013)
2. V. Schetinina, L. Jakaite, W. Krzanowski, Bayesian averaging over decision tree models: An application for estimating uncertainty in trauma severity scoring. *Int. J. Med. Inform.* **112**, 6–14 (2018)
3. V. Schetinina, L. Jakaite, W. Krzanowski, Bayesian averaging over decision tree models for trauma severity scoring. *Artif. Intell. Med.* **84**, 139–145 (2018)
4. V. Schetinina, L. Jakaite, W. Krzanowski, Bayesian learning of models for estimating uncertainty in alert systems: Application to air traffic conflict avoidance. *Integr. Comput. Aided Eng.* **25**(3), 229–245 (2018)
5. N. Selitskaya, S. Sielicki, L. Jakaite, V. Schetinina, F. Evans, M. Conrad, P. Sant, Deep learning for biometric face recognition: Experimental study on benchmark data sets, in *Deep Biometrics*, (Springer, Cham, 2020), pp. 71–97
6. J. Uglov, L. Jakaite, V. Schetinina, C. Maple, Comparing robustness of pairwise and multiclass neural-network systems for face recognition. *EURASIP J. Adv. Signal Process.* **2008**(1), 468693 (2007)
7. N. Nyah, L. Jakaite, V. Schetinina, P. Sant, A. Aggoun, Evolving polynomial neural networks for detecting abnormal patterns, in *IEEE 8th International Conference on Intelligent Systems (IS)*, (IEEE, Piscataway, 2016), pp. 74–80
8. N. Nyah, L. Jakaite, V. Schetinina, P. Sant, A. Aggoun, Learning polynomial neural networks of a near-optimal connectivity for detecting abnormal patterns in biometric data, in *2016 SAI Computing Conference*, (IEEE, Piscataway, 2016), pp. 409–413
9. M. Akter, L. Jakaite, Extraction of texture features from x-ray images: Case of osteoarthritis detection, in *Third Inter-*

- tional Congress on Information and Communication Technology*, (Springer, Singapore, 2019), pp. 143–150
10. L. Jakaite, V. Schetinina, C. Maple, Bayesian assessment of newborn brain maturity from two-channel sleep electroencephalograms, in *Computational and Mathematical Methods in Medicine* (2012), pp. 1–7
 11. L. Jakaite, V. Schetinina, C. Maple, J. Schult, Bayesian decision trees for EEG assessment of newborn brain maturity, in *The 10th Annual Workshop on Computational Intelligence UKCI 2010* (2010)
 12. L. Jakaite, V. Schetinina, J. Schult, Feature extraction from electroencephalograms for Bayesian assessment of newborn brain maturity, in *24th IEEE International Symposium on Computer-Based Medical Systems* (2011)
 13. G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. E. Learned-Miller and A. Ferencz and F. Jurie*, Marseille, France (2008)
 14. A. Dantcheva, C. Chen, A. Ross, Can facial cosmetics affect the matching accuracy of face recognition systems? in *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (Sept 2012), pp. 391–398
 15. R. Feng, B. Prabhakaran, Facilitating fashion camouflage art, in *Proceedings of the 21st ACM International Conference on Multimedia, MM '13*, (ACM, New York, 2013), pp. 793–802
 16. R. Min, A. Hadid, J.L. Dugelay, Improving the recognition of faces occluded by facial accessories, in *FG 2011, 9th IEEE Conference on Automatic Face and Gesture Recognition*, March 21–25, 2011, Santa Barbara, CA, USA
 17. C. Chen, A. Dantcheva, T. Swearingen, A. Ross, Spoofing faces using makeup: An investigative study, in *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)* (Feb 2017), pp. 1–8
 18. M. Eckert, N. Kose, J. Dugelay, Facial cosmetics database and impact analysis on automatic face recognition, in *IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)* (Sept 2013), pp. 434–439
 19. S. Mitra, M.I. Gofman, G. Parsons, J. Peissig, Facial asymmetry versus facial makeup, in *IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (2018), pp. 197–203
 20. S. Wang, Y. Fu, Face behind makeup, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16*, (AAAI Press, Palo Alto, 2016), pp. 58–64
 21. C. Chen, A. Dantcheva, A. Ross, Automatic facial makeup detection with application in face recognition, in *International Conference on Biometrics (ICB)* (2013), pp. 1–8
 22. CyberExtruder. Face matching data set biometric data [Online], <https://cyberextruder.com/face-matching-data-set-download>. Accessed 12 Aug 2019
 23. V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, R. Chellappa, Disguised faces in the wild, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2018), pp. 1–18
 24. A. Martinez, R. Benavente, The AR face database. Tech. Rep. 24, Computer Vision Center, Bellaterra (Jun 1998)
 25. S. Setty, M. Husain, et al., Indian movie face database: A benchmark for face recognition under wide variations, in *Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics* (2013), pp. 1–5
 26. A. Colombo, C. Cusano, R. Schettini, UMB-DB: A database of partially occluded 3D faces, in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (Nov 2011), pp. 2113–2119
 27. S. Selitskiy, N. Selitskaya, M. Koloeridi, BookClub artistic makeup and occlusions face data. Mendeley Data 2 (Sept 2020), <https://data.mendeley.com/datasets/yfx9h649wz/2>
 28. N. Selitskaya, S. Sielicki, N. Christou, Challenges in face recognition using machine learning algorithms: Case of makeup and occlusions, in *Advances in Intelligent Systems and Computing*, ed. by K. Arai, S. Kapoor, R. Bhatia, (Springer, Cham, 2020)
 29. S. Selitskiy, N. Christou, N. Selitskaya, Code for paper “Benchmarking accuracy and precision of the convolutional neural networks for face recognition on makeup and occluded images” (Jan 2021), <https://github.com/Selitskiy/ITNG2021>

Flávio Mota, Melise Paula, and Isabela Drummond

Abstract

The popularization of social networks has considerably increased the volume of data generated from the interaction between people. Understanding this data can be useful both for companies and governments and for users. This work proposes to study how to infer the behavior of people on social networks from published comments, specifically using the Myers-Briggs Typological Indicator (MBTI) in a social network focused on discussions on behavioral issues. The analysis carried out employs Natural Language Processing (NLP) techniques, resampling of the data set and classification algorithms combined by Majority Vote. The results showed 90% efficiency of the combiner with the use of random oversampling. SVM and KNN were the best individual classifiers regardless of the resampling technique used. Although smaller compared to the best individual classifier, the combination approach shows a decrease in the misclassification for INFJ and INFP classes up to 11% and 34%, respectively.

Keywords

Ensemble classifiers · Machine learning · Natural language processing · Resampling · MBTI · Personality · SMOTE · Majority vote · Social network · Behavior profile

F. Mota (✉) · M. Paula · I. Drummond
Institute of Mathematics and Computation Federal University of Itajubá, Itajubá, Brazil
e-mail: flaviomota@unifei.edu.br; melise@unifei.edu.br;
isadrummond@unifei.edu.br

59.1 Introduction

Content sharing on social networks provides a variety of information that makes it possible to discover users' experiences, opinions and feelings [1]. One way to discover knowledge in these environments is through text mining. According to [2] text mining can be considered a set of methods that are used to discover patterns in unstructured data.

Pattern extraction refers to the application of machine learning algorithms to process and find patterns from pre-processed data. A class of these algorithms is known as classifiers, which are used in predictive tasks [3].

In order to obtain better results in the classification task, it is possible to combine different classifiers. This combination, also known as ensemble, aims to bring together the predictions made by individual classification algorithms, resulting in a classifier that is generally more accurate than the individual ones [4].

The classification task can be affected by unbalanced data sets. When dealing with real-world problems such as text mining, some classes in the set may not be as well represented in relation to others [5]. Thus, resampling techniques can be applied to the set, to improve the accuracy.

In this context, this work focuses on combining classifiers to improve the classification results in some sense. It is known that, even if the classifier accuracy is not better, we can obtain better results considering each class, which can represent significant results in real applications. We proceed to investigate how the combined classifiers perform using different resampling techniques. The data set, comes from a social network, consists of comments that express people's behavior and opinions. The analysis is based on behavior patterns, such as those defined in the Myers-Briggs theory [6].

Table 59.1 The 16 personalities of the MBTI

Acronym	Typology
INTJ	Introversion intuition thinking judgment
INTP	Introversion intuition thinking perception
INFJ	Introversion intuition feeling judgment
INFP	Introversion intuition feeling perception
ISTJ	Introversion sensation thinking judgment
ISTP	Introversion sensation thinking perception
ISFJ	Introversion sensation feeling judgment
ISFP	Introversion sensation feeling perception
ENTJ	Extroversion intuition thinking judgment
ENTP	Extroversion intuition thinking perception
ENFJ	Extroversion intuition feeling judgment
ENFP	Extroversion intuition feeling perception
ESTJ	Extroversion sensation thinking judgment
ESTP	Extroversion sensation thinking perception
ESFJ	Extroversion sensation feeling judgment
ESFP	Extroversion sensation feeling perception

This paper is organized as follows. Section 59.2 presents the main concepts used in this research and Sect. 59.3 describes some related works. In Sect. 59.4 we show the database and work methodology. Section 59.5 discusses the results obtained and, finally, the conclusions and future work are presented in Sect. 59.6.

59.2 Background

This work aims to study the identification of classes related to Myers-Briggs Typology Indicator (MBTI) and the application of the combination by Majority Vote. Sections 59.2.1 and 59.2.2 present, respectively, a brief contextualization regarding the indicator and the technique employed.

59.2.1 Myers-Briggs Typology Indicator

Myers-Briggs Typology Indicator (MBTI) is a psychometric scheme proposed in 1950 by Katherine Cook Briggs and her daughter Isabel Briggs Myers, inspired by Freud and Jung's theories of personalities [6]. The scheme divides personalities into four axes and between motivations/preferences pairs: Extroversion/Introversion (E-I), Intuition/Sensation (N-S), Feeling/Thinking (F-T), and Judgment/Perception (J-P). The crossing of these 4 axes produces 16 psychological types, as shown in Table 59.1.

Each pair in the scheme has personality characteristics that are dichotomous. The first axis, E-I, is related to the individual's disposition. Extroverted individuals direct their psychic energy towards people and events in the external

environment. Introverts direct their attention to thoughts and experiences of the internal environment [6].

The N-S axis is related to the perception functions. According to [7], functions are a form of psychic activity that in principle remains unchanged even under varying conditions. Perceptual functions refer to how the individual acquires information, which can be intuitive(N) or sensory(S).

Judging functions are related to the F-T axis and refer to the way the individual draws conclusions about what is perceived, in this case through emotions(F) or rationalizing(T) [8].

Finally, the J-P axis refers to the orientation mode in relation to the external world. Judging(J) individuals have an organized and planned lifestyle, evaluating and controlling data from reality, which implies structured and methodical ways of dealing with the external world. Perceptual(P) individuals are more flexible and spontaneous, dealing with data from reality through openness to experiences. It implies creative and adaptive ways of dealing with the outside world [9].

There are currently systems¹ that provide a person's profile by applying a questionnaire. These profiles are used in corporate team building exercises, student assessment for educational planning, among others [10].

59.2.2 Combining Classifiers

According to [11] the combination of classifiers is the application of several classifiers that have their results aggregated to produce an output that determines the class of an object. This combination tends to produce more accurate results than a single classifier.

Majority vote is a way of combining classifiers. In this approach the class resulting from the combination is determined by the class that was found by most classifiers [12]. Figure 59.1 demonstrates the voting approach.

The data set is used to train m classifiers (C_1, C_2, \dots, C_m), whose predictions (P_1, P_2, \dots, P_m) are combined through voting, and the final prediction P_f is the class that received the most votes.

59.3 Related Works

Liu et al. [13] use the Naïve Bayes algorithm to analyze posts in personal microblogs made available in an API and determine the user's profile according to the Myers-Briggs behavioral theory. Applying a process of collection, pre-processing, representation, selection of attributes and training, the authors evaluated the performance of the algorithm

¹16personalities – <https://www.16personalities.com/>

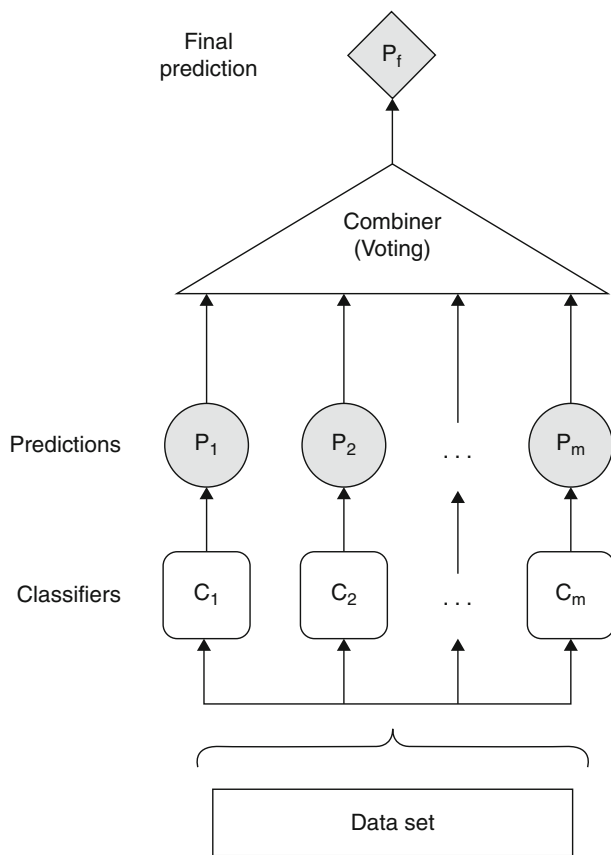


Fig. 59.1 Voting combination approach

in texts written in Chinese. The model was trained and tested using the API data and the results show the effectiveness of the algorithm applied to the theory.

The work of [10] analyzes feature extraction techniques in a database collected from a social network called Personality Cafe.² The database contains 8675 records, with 50 posts from each user and their typological indicator present in the Myers-Briggs theory. The authors computed the Category of Term (CAT), Term Nuance, Terms Frequency (TF), Inverse Document Frequency (IDF) and Term Frequency – Inverse Document Frequency (TFIDF) to analyze, through of distance metrics, how much each technique can generate a representation for the set. The results show that the Category of Term generated better representation values for the data set classes.

We can see in [12] an analysis of techniques for combining classifiers for text mining. Using data from the ACM repository, the authors carried out an extensive empirical study of the Naïve Bayes, SVM, Logistic Regression and Random Forest algorithms and the AdaBoost, Bagging, Dagging, Random Subspace and Majority Vote techniques to identify

Table 59.2 Data set example

Type	Posts
INTJ	Dear INTP, I enjoyed our conversation the other day. Esoteric gabbing about the nature of the universe and the idea that every rule and social code being arbitrary constructs created... Dear ENT...
ENTJ	You're fired. That's another silly misconception. That approaching is logically is going to be the key to unlocking whatever it is you think you are entitled to. Nobody wants to be approached wit...
INTP	I'm in this position where I have to actually let go the person, due to a various reason. Unfortunately I'm having trouble mustering enough strength ...

keywords in the base. The results demonstrate that, for the data set employed, the use of the Random Forest with the Bagging technique provided the best results.

59.4 Material and Methods

The data collected for this work come from a discussion forum on the Myers-Briggs theory of the Personality Cafe network. In this forum, users with different typological profiles interact on different subjects from their profile point of view. The database contains 8675 records with the acronym identifying the user's type and a collection of 50 publications (posts) separated by the string "|||". An example of the data can be seen in Table 59.2.

59.4.1 Pre-processing

In this step, the data were pre-processed to allow the classification algorithms to be applied. All links, special characters and punctuation have been removed from the text that makes up the publications, and references to other users have been replaced by the word *user*. The text has also been converted to lowercase.

After cleaning the set, we use the bag-of-words technique to represent textual data numerically. For this, we use two approaches: Term Frequency (TF) and Term Frequency – Inverse Document Frequency (TFIDF). For the application of the TF, we consider stop-words *and, the, to, of, infj, entp, intp, intj, entj, enfj, infp, enfp, isfp, istp, isfj, istj, estp, esfp, estj, esfj, infjs, entps, intps, intjs, entjs, enfjs, infps, enfps, isfps, istps, isfjs, istjs, estps, esfjs, estjs, esfjs*, as these are terms that appear in the documents but do not contain relevant information.

The occurrence of the same word in many documents may indicate that it does not add relevant information for data analysis. Therefore, the TFIDF technique is applied to decrease the values of the vectors constructed in the previous step [14].

²Personality Cafe – <https://www.personalitycafe.com/>

59.4.2 Resampling

When analyzing the distribution of the class occurrences, it is possible to observe that the set is unbalanced, containing many more occurrences of the INFP class than ESTJ, as can be seen in Fig. 59.2. According to [5] unbalanced data sets cause problems for classifiers since there are not enough data from minority classes for models to learn.

There are different approaches in the literature regarding data resampling [15]. These approaches are commonly divided into two types: subsampling, in which instances of the majority classes are removed, and oversampling, in which new instances of the minority classes are generated. There are also hybrid approaches that combine subsampling and oversampling [16]. In this work, we employ oversampling techniques and a hybrid technique in order to allow a comparison of the results obtained with the study data set.

We consider two oversampling techniques: the random oversampling and the SMOTE technique (Synthetic Minority Over-sampling Technique). The first one consists of generating new instances of minority classes by randomly replicating the original instances [17]. And the SMOTE technique generates synthetic instances between the lines that join an instance of a minority class to its k closest neighbors of the same minority class [18].

The hybrid technique used, presented in [19], combines the application of the SMOTE technique and of Tomek links. Tomek links are pairs of instances x and y that belong to different classes with a distance $d(x, y)$ for which there is no instance z such that $d(x, z) < d(x, y)$ or $d(y, z) < d(y, x)$. Given that the SMOTE technique creates artificial instances of minority classes that can be understood as noise, the hybrid technique aims to remove from the set both instances of minority classes and instances of majority classes that form a Tomek link.

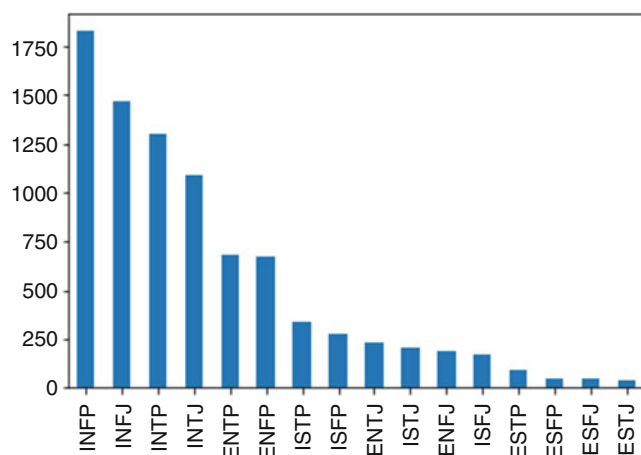


Fig. 59.2 Distribution of classes in the set

To apply the chosen techniques, we use the Python language library *imbalanced-learn*, with the implementations *RandomOverSampler* for random oversampling, *SMOTE* for SMOTE and *SMOTETomek* for the hybrid technique SMOTE with Tomek links.

59.4.3 Classification

Based on the results of Refs. [2, 20–22], we chose four (4) classification techniques: Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and K-next neighbor (KNN). The technique selected to combine them is the Majority Vote (MV). And, from Python language, we employ the implementations from *scikit-learn* library.

The library provides several versions of the SVM algorithm, and in this work we use the *SVC* version, considering as parameters the linear *kernel* and regularization (C) of 100. For the Naïve Bayes algorithm, we chose the *MultinomialNB* implementation, the one recommended by the developers as suitable for text mining. For the DT, we use the standard (*DecisionTreeClassifier*), and the KNN (*KNeighborsClassifier*) algorithm consider parameters k equal to 1 and the Euclidean distance metric ($p = 2$). Finally, the combination was implemented with the *VotingClassifier*, taking as parameters the previously mentioned classifiers and majority voting style of predicted labels (*voting='hard'*).

To perform cross-validation we use function *cross validate*, setting the parameter n_jobs to -1 , which means that all CPUs of the machine are used during the training and testing process. The function returns the accuracy values of the classifiers. The experiments were performed on a machine with a *Intel®Core™i5-9600K CPU @ 3.70GHz* processor with 6 cores and 8GB of RAM.

59.5 Results

The evaluation of the results obtained considered the metrics of accuracy, F1, precision and recall using cross-validation with five folders, in addition to the execution time.

Tables 59.3, 59.4, and 59.5 present the values obtained with the application of the classifiers for the different resampling techniques employed.

In general, regardless of the resampling technique, the SVM classifier obtained the best results, with an accuracy of approximately 90%, being slightly better than the combiner in all cases. The *Naïve Bayes* model obtained the worst accuracy in all scenarios. The Decision Tree showed values of approximately 87% with random oversampling. On the other hand, applying the SMOTE and SMOTE + Tomek techniques, there was a drop in the classifier's hit rate, approaching the values obtained by *Naïve Bayes*. It is worth

Table 59.3 Results for random oversampling

Random oversampling					
	Accuracy	F1	Precision	Recall	Time (s)
SVM	0.91 ± 0.03	0.91 ± 0.03	0.91 ± 0.04	0.91 ± 0.03	876.75
NB	0.53 ± 0.01	0.53 ± 0.01	0.53 ± 0.01	0.53 ± 0.01	0.59
DT	0.87 ± 0.05	0.85 ± 0.04	0.85 ± 0.06	0.87 ± 0.05	20.35
KNN	0.89 ± 0.06	0.89 ± 0.06	0.93 ± 0.05	0.89 ± 0.06	86.33
MV	0.90 ± 0.04	0.89 ± 0.04	0.90 ± 0.04	0.90 ± 0.04	942.53

Table 59.4 Results for smote oversampling

SMOTE					
	Accuracy	F1	Precision	Recall	Time (s)
SVM	0.90 ± 0.04	0.90 ± 0.04	0.90 ± 0.04	0.90 ± 0.04	1180.60
NB	0.58 ± 0.04	0.58 ± 0.03	0.58 ± 0.03	0.58 ± 0.04	0.52
DT	0.59 ± 0.05	0.58 ± 0.03	0.59 ± 0.02	0.59 ± 0.05	13.43
KNN	0.87 ± 0.04	0.85 ± 0.05	0.85 ± 0.06	0.87 ± 0.04	153.75
MV	0.88 ± 0.04	0.87 ± 0.04	0.88 ± 0.04	0.88 ± 0.04	1335.82

Table 59.5 Results for resampling with SMOTE + Tomek

SMOTE + Tomek					
	Accuracy	F1	Precision	Recall	Time (s)
SVM	0.90 ± 0.04	0.90 ± 0.04	0.91 ± 0.04	0.90 ± 0.04	1199.32
NB	0.57 ± 0.04	0.57 ± 0.04	0.58 ± 0.03	0.57 ± 0.04	0.59
DT	0.59 ± 0.04	0.58 ± 0.03	0.59 ± 0.04	0.59 ± 0.04	14.04
KNN	0.87 ± 0.04	0.84 ± 0.05	0.85 ± 0.06	0.87 ± 0.04	206.00
MV	0.88 ± 0.04	0.88 ± 0.04	0.88 ± 0.04	0.88 ± 0.04	1409.57

noting that the oversampling techniques are different when they generate new data, which can be similar instances to the original (random oversampling), or the creation of synthetic samples based on the distances between the neighbors of the classes. This can justify the different results found, since the algorithms manipulate different values.

In proportion to its performance, the SVM execution time is the longest. In this respect, KNN presents results comparable to SVM in less time. Despite being the fastest, *Naïve Bayes* does not present satisfactory results in relation to the other algorithms. The execution time of the combination of classifiers is the longest, which is expected, since it uses the prediction made by all individual classifiers.

Analyzing the performance of the combiner, in general, no better results were obtained than the best individual classifier (SVM). This may be related to the fact that the majority voting strategy is being negatively influenced by the values obtained by the individual classifiers with the worst results. However, by performing an analysis between the two best classifiers (SVM and KNN), it was possible to notice that, in all scenarios, the combiner reduced the classification error of the INFJ and INFP classes. Through random oversampling, the error reduction of the INFJ class was 11% in relation to the KNN and of the INFP class was of 5% in relation to the

SVM. With the SMOTE technique, the INFJ class had its error reduced by 6% and the INFP class by 34% in relation to the KNN. The ENFP class had the error reduced by 4% in relation to the classification made by the SVM. For the scenario with resampling with SMOTE + Tomek, in relation to the KNN, the classes INFJ, INFP and INTP obtained an error reduction of 5%, 33% and 5% respectively. In this scenario, the ENFP class achieved a 4% error reduction compared to the SVM. The INFP and INFJ classes are the two major classes in the set and, according to behavioral theory, they are distinguished only by the last axis (Judgment/Perception), which may justify the difficulty found by the classifiers in separating them. The ENTP and ISFJ classes, considered as dichotomous in relation to behavioral theory, were classified without error by all classifiers and in all scenarios.

59.6 Conclusions and Future Work

In this work we have combining classifiers to identify behavioral profiles, using the Myers-Briggs typological indicator, also considering the application of different resampling techniques in the data set. The results show that the combination approach by Majority Vote got 90% correct with the random oversampling and 88% with the SMOTE and SMOTE + Tomek techniques. Even so, it was not better than the best among the individual classifiers (SVM), although in certain scenarios we can observe a reduction of up to 11% and 34% in the classification error of the INFP and INFJ personalities, respectively. These two classes differ only by the last axis of personality (Judgment/Perception), which makes them very similar. Thus, the combination of classifiers was able to improve the classification, reducing the error of the individual classifiers in these classes. Only the NB classifier was not adequate in relation to the data set in any scenario. The DT model had its performance reduced when applying resampling with SMOTE and SMOTE + Tomek.

We can mention as possible future works the use of other techniques for combining classifiers, such as Bagging and AdaBoost, as well as the use of other classification models. Also, the use of pre-processing techniques such as POS (Part-Of-Speech) or information gain can be investigated. In addition, this work considered the application of oversampling approaches, the subsampling must be tested and its performance compared in future works.

References

1. E. Souza, D. Vitória, D. Castro, A.L.I. Oliveira, C. Gusmão, Characterizing opinion mining: A systematic mapping study of the Portuguese language, in *Computational Processing of the Portuguese Language, Portugal*, vol. 9727, (Springer, Cham, 2016), pp. 122–127

2. R.A. Sinoara, J. Antunes, S.O. Rezende, Text mining and semantics: A systematic mapping study. *J. Braz. Comput. Soc.* **23**, 9 (2017)
3. K. Faceli, A.C. Lorena, J. Gama, A.C.P.L.F. de Carvalho, *Inteligência artificial: Uma abordagem de aprendizado de máquina* (Editora LTC, Rio de Janeiro, 2011)
4. D. Opitz, R. Maclin, Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999)
5. C. Padurariu, M.E. Breaban, Dealing with data imbalance in text classification. *Procedia Comput. Sci.*, Budapest, Hungary **159**, 736–745 (2019)
6. I.B. Myers, P.B. Myers, *Gifts Differing: Understanding Personality Type*, 2nd edn. (Davies-Black, Boston, 2010)
7. C.G. Jung, *O eu e o inconsciente*, 27th edn. (Editora Vozes, Petrópolis, 2015)
8. G. Couto, D. Bartholomeu, J.M. Montiel, Estrutura interna do Myers Briggs Type Indicator (MBTI): evidência de validade. *Avaliação Psicológica* **15**(1), 41–48 (2016)
9. I.B. Myers, L.K. Kirby, K.D. Myers, *Introduction to Type: A Guide to Understanding Your Results on the Myers-Briggs Type Indicator*, 6th edn. (Consulting Psychologists Press, Palo Alto, 2000)
10. C. Li et al., Feature extraction from social media posts for psychometric typing of participants, in *AC 2018: Augmented Cognition: Intelligent Technologies, Las Vegas, NV, USA*, vol. 10915, (Springer, Cham, 2018), pp. 267–286
11. L.I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd edn. (Wiley, Hoboken, 2014)
12. A. Onan, S. Korukoğlu, H. Bulut, Ensemble of keyword extraction methods and classifiers in text classification. *Expert Syst. Appl.* **57**, 232–247 (2016)
13. Y. Liu, T. Liu, Y.J. Wang, Research on micro-blog character analysis based on Naïve Bayes, in *Seventh International Conference on Digital Image Processing (ICDIP 2015), Los Angeles, United States*, vol. 9631, (SPIE, Bellingham, Washington, 2015), pp. 549–553
14. F. Provost, T. Fawcett, *Data Science Para Negócios*, 1st edn. (Alta Books, Rio de Janeiro, 2016)
15. N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study. *Intell. Data Anal.* **6**(5), 429–449 (2002)
16. N. Junsomboon, T. Phientrakul, Combining over-sampling and under-sampling techniques for imbalance dataset, in *Proceedings of the 9th International Conference on Machine Learning and Computing*, Singapore (2017), pp. 243–247
17. T. Zhu, Y. Lin, Y. Liu, Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recogn.* **72**, 327–340 (2017). <https://doi.org/10.1016/j.patcog.2017.07.024>
18. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**(1), 321–357 (2002)
19. G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **6**(1), 20–29 (2004)
20. R.E.S. Santos, J.S. Correia-Neto, E.P.R. Souza, C.V.C. de Magalhães, G. Vilar, Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: Resultados preliminares de um mapeamento sistemático. *Revista de Sistemas e Computação* **4**(2), 116–125 (2014)
21. E. Souza, D. Vitória, D. Castro, A.L.I. Oliveira, C. Gusmão, Characterizing opinion mining: A systematic mapping study of the Portuguese language, in *Computational Processing of the Portuguese Language, Portugal*, vol. 9727, (Springer, Cham, 2016), pp. 122–127
22. H. Soong, N.B.A. Jalil, R.K. Ayyasamy, R. Akbar, The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and techniques, in *2019 IEEE 9th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, Kota Kinabalu, Malaysia (2019), pp. 272–277

Sunil Manzoor and Shahram Latifi

Abstract

The purpose of this study is to develop a model which employs the facial expressions and features of people to predict their health. Our objective is to find the best Machine learning approaches to develop a health model which utilizes the facial features. This report also discusses the available datasets of facial expressions. Here, we utilize such machine learning techniques as regression, neural network, and clustering to predict symptoms of sickness. To construct the model, we train our model with the healthy people images acquired from JAFFE database. After that, we ran the test dataset that includes an equal amount of sick and healthy people images. Utilizing the CCN (convolutional neural network) approach, our model has been able to predict the health of a person based on the facial features with an accuracy of 70%. This model could be utilized as the first level of diagnosis and can be implemented to distinguish between a healthy and sick person at the entrance of the public facilities. Such information could be crucial in the prevention and control of infectious diseases.

Keywords

Computer vision · Computer-aided · Disease detection · Facial features · Facial expressions · Image classification · Image segmentation · Neural network · Preliminary diagnosis · Temporal variations

S. Manzoor · S. Latifi (✉)
Electrical and Computer Engineering, University of Nevada, Las Vegas, NV, USA
e-mail: manzoor@unlv.neveda.edu; shahram.latifi@unlv.edu

60.1 Introduction

Machine Learning (ML) and Artificial Intelligence (AI) are emerging new disciplines in computer science with their applications expanding to all major health care departments including medical image processing, computer-aided diagnosis, image interpretation, image fusion, image registration, image segmentation, image-guided therapy, image retrieval, and analysis [1].

Recently, computer vision has become a powerful tool, especially in the medical field. X-Ray and MRI images are being analyzed through the computer so doctors can diagnose with more efficiency and accuracy. Dentists are also using many machine learning-based solutions for diagnostic purposes. However for all the above mentioned uses facial expressions are the most important factors to detect what a person is going through. Rahu et al. [2] recorded expressions of people who were non communicative critically ill people and showed how their expressions reflect pain during endotracheal suctioning. Multiple facial actions were included such as brow lowering and raising, nose wrinkles, eye closure, and others [2]. That shows how the inner situation of the body can be determined from the face. Another report by Wang et al. in 2016 identified the visually observable disease symptoms by observing common facial features [3] which are very nearly the same among people regardless of age, gender, and race. For this purpose, they used above 8200 images of different people with the front of their faces towards the camera. They used statistical methods to detect outliers that were different from the training data. Most of the work was based on finding visually observable diseases including disposition of eyeball, disoriented nose or eyes, and out of shapes lips. Therefore, they investigate facial expressions by using the combination of the perceptual and cognitive

sciences in addition to affective computing and computer animations [4]. That is why facial expression recognition has become cutting edge technology in the field of artificial intelligence and virtual reality. Therefore after researching facial expression recognition techniques we have come up with a model that is mainly focused on distinguishing between healthy and sick persons by processing the facial images of people. This is especially helpful in the current pandemic as people displaying flu-like facial symptoms can be diagnosed early as sick as the human face plays a prodigious role for automatic recognition of expressions [5].

In our model, we extract features such as eyes, nose, and upper lips from normal people images dataset with computer vision techniques such as active shape models (ASMs) and compare them with our extracted features from image collection of sick people who display flu symptoms. The imagery data collected include both genders with age between 25 to 70 years. In Sect. 60.2 we have discussed the common approaches that currently being used and in Sect. 60.3 we have discussed available facial images datasets and their characteristics. Method and flow chart of the prescribed model is described in Sect. 60.4. In Sect. 60.5 the results are shown and analysis is done before going to conclusion that is in Sect. 60.6.

60.2 Facial Expressions Recognition Approaches

Extracting facial features through classification is not a recent trend people have already been trying different methods to extract features since the 1960s [6]. Most of the earlier such reports when computer vision was a relatively new term were primarily based on methods such as linear discriminant analysis, principal component analysis, and independent component analysis to extract facial features until Ferry et al. [7, 8] came up with the idea of SVM classifier to extract features in 2014.

Face recognition technology has matured gradually as compared to fingerprint recognition, and iris recognition because face accuracy can easily be affected by factors such as light present and the expressions when the picture was taken [9]. Therefore it is really important to keep these factors in mind while extracting the features.

Expression detection through facial feature extraction is an interesting practice people are working on. A recent report by Dagar et al. [10] have designed an automated framework that detects human expressions based on their facial expression in 2016. Therefore, in this report, we have researched some techniques currently being used to detect features from pictures and found two major categories conventional and deep learning-based methods.

In conventional ways, three are the main steps that every algorithm is based on. The first of which is detecting the major components of a face like eyes, nose, lips etc. After that extracting features from components and classifying those features is the third step. In contrast, modern approaches use deep learning algorithms [4].

Most of the conventional approaches used for geometric features would investigate a relationship among features to make a feature vector. The next step is to locate landmarks from that feature vector before coming up with angles among them. For example, the right distance of the eyes from the nose and whether they are even with each other. The other is the appearance features that are to be taken from the entire face. Here we can use simple machine learning techniques like PCA and LBP [4]. Appearance features are the main target areas of the face, for example, we focus on the eyes more than the forehead if we want to examine the health of a person but if we want to explore skin color then we need to detect that from the forehead instead of eyes. Sometimes we use both the above-mentioned techniques to make it a hybrid features extraction approach.

Deep-Learning Based facial expression approaches utilize convolutional neural network (CNN) and recurrent neural network (RNN) in deep learning calculations that can be directly applied to the field of ML. CNN contains three types of heterogeneous layers convolution layer, max-pooling layer, and fully connected layers [4].

CNNs takes images and relates to them with a feature map then assigns different weights to the related features and ultimately decides something about the input based on assigned weights. Many approaches have evolved from CNNs but they cannot reflect temporal variations in the facial components. That is why some people have combined CNN with conventional approaches and produced good results. Kahou et al. [12] proposed a model with the combination of both RNN and CNN approaches to satisfy the restrictions generated in CNN and RNN in 2015.

60.3 Datasets Available for Facial Expressions Recognition

60.3.1 Japanese Female Facial Expressions (JAFPE) [5]

The JAFPE database comprises 213 facial images of ten different female Japanese models with each containing seven emotions with six facial emotions and one neutral emotion. Each image was rated based on six emotional adjectives using 60 Japanese subjects. The resolution of all the images are 256×256 pixels.

60.3.2 The MUCT Landmarked Face Database [11]

This database contains a total of 3755 images with 5 cameras a, b, c, d, and e positions of facial images of 276 subjects with blue background. The resolution of each facial image is 480×640 pixels. This dataset provides diversity age and ethnicity.

60.3.3 Extended Cohn-Kanade Dataset [13]

It contains 593 video sequences of 123 people with the following resolutions (640×480 and 640×490) for gray-scale values with 8-bit precision.

60.3.4 Compound Emotion (CE) [14]

The CE comprises 5060 images matching 22 categories of emotions (basic and compound) for human subjects (230) among which there were 130 females, 100 males with a mean ages of 23.

60.3.5 Binghamton University 3D Facial Expression (BU-3DFE) [16]

It comprises data from about 100 adult subjects among which there are 56 females and 44 males, presenting 6 emotions. There are 25 3D facial emotion models of each subject along with a set of 83 facial landmarks that have been manually annotated per each model. The size and resolution of each image is 1040×1329 pixels.

60.3.6 Denver Intensity of Spontaneous Facial Action Database (DISFA) [15]

It is comprised of 130,000 stereo video frames of a total of 27 adult subjects with different ethnicities at the resolution of 1024×768 among which there are 12 females and 15 males.

60.3.7 Extended Yale B face [17]

This database comprises 16,128 images taken of the faces of 28 distinct subjects under a single light source. The images display 576 viewing angles and nine different poses for each of the 64 illumination conditions. The resolution of every image is 320×243 pixels.

60.3.8 The Karolinska Directed Emotional Face (KDEF) [18]

This database consists of 4900 images of facial expressions of 70 individuals with each image displaying seven different emotional expressions shot from five different angles. The resolution of each facial image is 562×762 pixels.

60.3.9 The MUG Facial Expression Database [20]

This is huge data set of image sequences of six basic facial expressions of 86 people 51 men and 35 women having resolution of 896×896 pixels. Image format is “jpg” and video format is “avi”. It consist of 38GB.

60.4 Methods

60.4.1 Training Dataset

We used two separate types of image data for this model including a healthy person’s frontal face and symptomatic frontal images. In a normal data set, we have 213 images downloaded from the JAFFE Database [5] consisting of 10 Japanese females with six expressions and one neutral pose; examples are shown in Fig. 60.1. We also acquired 751 images from MUCT [11] that consist of a total of 3755 images with 5 cameras a, b, c, d, and e. For this study, we used images from camera “a” for frontal face view only; examples are shown in Fig. 60.2. With both these datasets we also used MUG dataset [20] but we only selected 380 images of happy expression and put them in healthy category and 380 images of sad expression for sick category; examples are shown in Fig. 60.3.

We created the dataset for healthy people by accumulating 30 images displaying neutral and 31 images of happy expressions from the JAFFE database, along with 751 images from MUCT database with camera position “a”, thereby forming

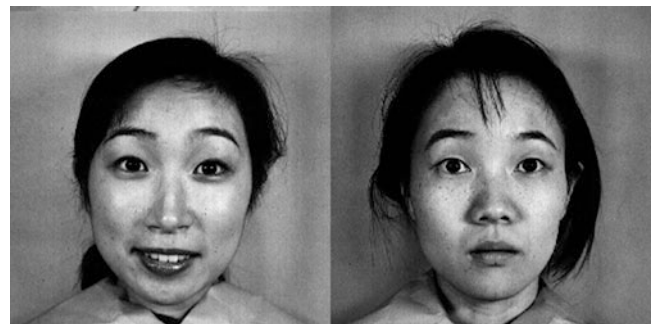


Fig. 60.1 JAFFE dataset [5]



Fig. 60.2 MUCT-A dataset [11]

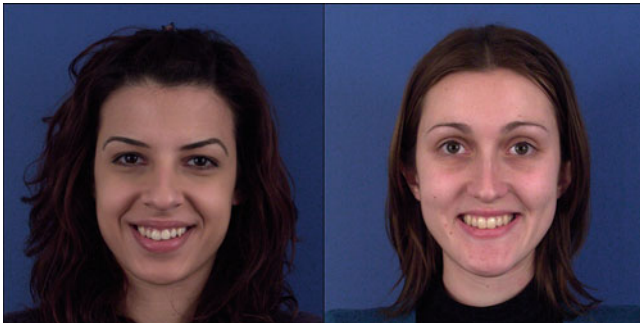


Fig. 60.3 MUG dataset [20]

a total of 812 images. Also, we used 760 images from the MUG dataset such as 380 images of smiling face and 380 images with sad expressions. To create a sick people dataset, we took 31 images displaying sad expressions as well as 29 images displaying an expression of disgust along with the 28 images downloaded from the internet displaying facial symptoms of flu, see Fig. 60.4. We used images of sadness and disgust as an indication of sickness as the inflammation or sickness can alter the emotional expressions and have a significant mediating role in disease detection as recently reported in a study conducted by Sarolidou et al. [19]. These are randomly selected images from different sources. These images of different sizes and resolutions so we manually made them the same size as originally present in the healthy individual dataset.

60.4.2 Extracting Features

As we used two different databases as mentioned above where JAFFE contains images with no RGB but MUCT has color images, we removed RGB values of all the images first and made them of equal size of 128×128 pixels. We provided the same treatment to randomly selected sick people



Fig. 60.4 Radom images dataset for sick people

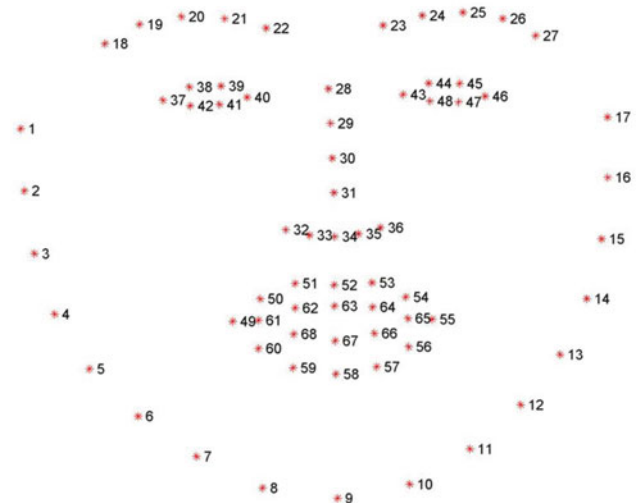


Fig. 60.5 Facial landmarks

images as well. After reading the images from the collection we took every single picture and marked it down with points around the face, then some points around both the eyes and nose. At the bottom of the nose, we mark points as well just to predict a runny nose for the disease. We call these points as labels, as seen in Fig. 60.5. These labels will form different virtual polygons over the face on the images. We observed while testing that sometimes these labels overlap each other, so to avoid that we kept only one entry of those labels, which removed the overlapping labels. This happens when we differentiate the area from upper-lips with nose or lips. We extracted around three to four different features from the images, including the nose, both eyes, the lips, and the area between upper lip and nose. If a picture had any problem detecting all the features we discarded it so it could not pollute our training model.

In order to detect that someone is sick, we provided our model with sick people images to train as well, and found people with features belonging to that class.

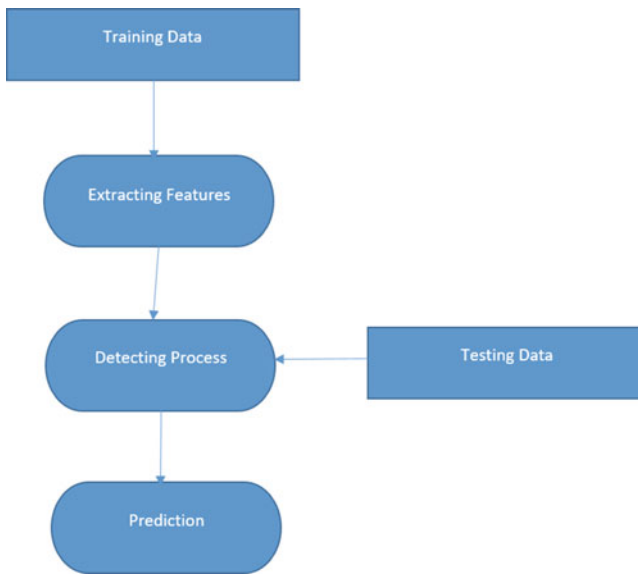


Fig. 60.6 Flow chart of the developed model

60.4.3 Detecting Process

We used a Convolutional Neural Network (CNN) with 15 layers and 30 loops to predict the results for randomly selected sample images from online resources. Our model extracted labels from the test sample image just like it did with the training dataset and compare the new image with its already learned extracted features.

60.4.4 Flow-Chart of Model (Fig. 60.6)

60.5 Results

In this study, we trained our model on overall 1660 images where 1192 for healthy class and 468 images in the category of sick faces. Eighty-five percent of the data were used to train the data and 15% of the data were used to test our model, using a neural network, eventually dividing the dataset into two labels: Sick and Healthy. After applying the CNN, we got 15 different layers with parameters. We got a total of 61,326 parameters, and all were trainable. We applied 30 iterations to fit our model and got the loss and accuracy, with their respective values at every epoch. On manual testing we gave 40 images of both categories and model predicted 28 images with the accuracy of 70%. We got the following loss and accuracy graphs after training the model, see Figs. 60.7 and 60.8.

As our model was trained on facial images contained in JAFFE, MUG and MUCT, so we gave images with face to get the best result. However, below are the two images predicted by our model, Fig. 60.9.



Fig. 60.7 training loss and validity loss

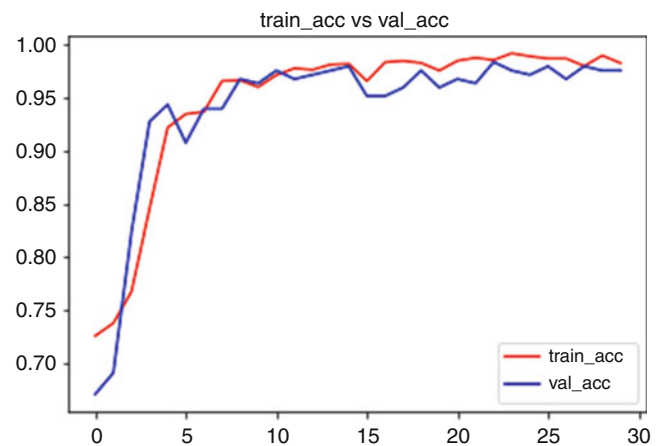


Fig. 60.8 training accuracy and validity accuracy



Fig. 60.9 The images projected by the model based on facial expression (a) A predicted image of sick person (b) a predicted image of a healthy person

60.6 Conclusion

This paper mainly focused on summarizing different methods of machine learning for facial feature recognition and listing the different databases for face images with their characteristics and limitations. It also came up with a model that has been trained on the facial images of healthy and sick

people and getting trained on them predicted sick people by comparing their extracted features with normal people.

In order to achieve our goal of detecting healthy and sick people, we used two different datasets of images of normal people, and trained our model on them as healthy people. We trained our model on another dataset as well that contained a real-world collection of face images of sick people. In some cases, where sick people had no visual symptoms, our models showed false-positive behavior.

Although many deep learning approaches have been utilized for Image processing especially in Healthcare sector, however, their penetrations in the Healthcare applications are a slow moving process. The results of our modeling indicate that the CNN method is slow but showed satisfactory results on testing but further studies are needed to improve the accuracy.

Acknowledgement We thank and appreciate the support of Meagan Madariaga-Hopkins for the assistance in editing this manuscript.

References

1. M.I. Razzak, S. Naz, A. Zaib, "Deep Learning for Medical Image Processing: Overview, Challenges and the Future." *Classification in BioApps* (Springer, Cham, 2018), pp. 323–350
2. M.A. Rahu, M.J. Grap, J.F. Cohn, C.L. Munro, D.E. Lyon, C.N. Sessler, Facial expression as an indicator of pain in critically ill intubated adults during endotracheal suctioning. *Am. J. Crit. Care* **22**(5), 412–422 (2013)
3. K. Wang, J. Luo, Detecting visually observable disease symptoms from faces. *EURASIP J. Bioinforma. Syst. Biol.* **2016**(1), 13 (2016)
4. B.C. Ko, A brief review of facial expression recognition based on visual information. *Sensors* **18**(2), 401 (2018)
5. M. Lyons, M. Kamachi, J. Gyoba, Japanese Female Facial Expression (JAFFE) Database. figshare. Journal contribution. <https://doi.org/10.6084/m9.figshare.5245003.v2> S (2017)
6. W.W. Bledsoe, *The model method in facial recognition*, vol 15 (Panoramic Research Inc, Palo Alto, 1966), p. 47
7. Q. Ferry, J. Steinberg, C. Webber, D.R. FitzPatrick, C.P. Ponting, A. Zisserman, C. Nelläker, Diagnostically relevant facial gestalt information from ordinary photos. *elife* **3**, e02020 (2014)
8. X.-Y. Li, Z.-X. Lin, The Euro-China Conference on Intelligent Data Analysis and Applications, in *Face Recognition Based on HOG and Fast PCA Algorithm*, (Springer, Cham), p. 2017
9. P. Wei, Z. Zhou, L. Li, et al., Research on face feature extraction based on K-mean algorithm. *J. Image Video Proc.* **2018**, 83 (2018). <https://doi.org/10.1186/s13640-018-0313-7>
10. D. Dagar, A. Hudait, H.K. Tripathy, M.N. Das, *Automatic expression detection model from facial expression*. In *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pp. 77–85. IEEE (2016, May)
11. J.M. Milborrow, F. Nicolls. The MUCT Landmarked Face Database. In *Proc. Pattern Recognition Association of South Africa*, 2010
12. S.E. Kahou, V. Michalski, K. Konda. Recurrent neural networks for emotion recognition in video. In *Proceedings of the ACM on International Conference on Multimodal Interaction*, Seattle, WA, USA, 9–13 November 2015, pp. 467–474
13. P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, CA, USA, 13–18 June 2010, pp. 94–101
14. S.Y. Tao, A.M. Martinez, Compound facial expressions of emotion. *Natl. Acad. Sci.* **111**, E1454–E1462 (2014)
15. S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J. Cohn, DISFA: a spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* **4**, 151–160 (2013)
16. L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3D facial Expression database for facial behavior research. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, 10–12 April 2006, pp. 211–216
17. B+. Available online: <https://computervisiononline.com/dataset/1105138686> (accessed on 29 November 2017)
18. KDEF. Available online: <http://www.emotionlab.se/resources/kdef> (accessed on 3 November 2020)
19. G. Sarolidou, J. Axelsson, T. Sundelin, J. Lasselin, C. Regenbogen, K. Sorjonen, et al., Emotional expressions of the sick face. *Brain Behav. Immun.* **80**, 286–291 (2019)
20. N. Aifanti, C. Papachristou, A. Delopoulos, The MUG facial expression database. 1–4 (2010)

Performance Comparison of Algorithms Involving Automatic Learned Features and Hand-Crafted Features in Computer Vision

Rocky Y. Gonzalez and Shahram Latifi

Abstract

Three case studies in this report provide an analysis between two different machine learning algorithms in computer vision (CV) systems, with the difference in automatic feature learning as found in Deep Learning (DL), and manual hand-crafting feature extraction/selection, as found in traditional CV methods. Furthermore, this report focuses on state-of-the-art between Convolution Neural Networks (CNNs) and Scale-Invariant Feature Transform (SIFT) related-models.

In the first case study, a CNN algorithm is trained by a set of 8000 images to binary classify 2000 test images. Adding more epochs and layers to the CNN indicates an increase of performance for the training set and testing set with a trade-off in increased system complexity. In the second case study, an image retrieval system is implemented using SIFT, SIFT-filtered, and CNN algorithms to evaluate an Oxford-5 k dataset. Results using a baseline SIFT and SIFT-filtered algorithm show the mean average precision (mAP) performance to be better than that performed by the CNN algorithm. In the third case study, different datasets are evaluated for image instance retrieval using the CNN and the Fisher Vector (FV) involving SIFT descriptors. Depending on the dataset being evaluated, a baseline FV shows the mAP to be comparable in performance to that of the CNN.

Keywords

Binary classification · Convolutional neural network (CNN) · Deep learning (DL) · Feature descriptor · Feature filtering · Fisher vector (FV) · Image retrieval ·

Mean average precision (mAP) · Scale-invariant feature transform (SIFT) · Traditional computer vision (CV)

61.1 Introduction

Computer Vision (CV) is a field of study in Machine Learning (ML) that seeks to extract information from an image or images [1]. Some of the different applications that fall under the field of computer vision include Optical Character Recognition (OCR), machine inspection, automated check-out lanes (in retail), 3d model building, medical imaging, self-driving vehicles, motion capture, biometrics, and more [1]. With technological improvements in computer GPU processes, and the emergence of large-scale labelled data, Deep Learning (DL) algorithms have become increasingly more common in solving problems within machine learning in general [2, 3].

This research effort seeks to distinguish the differences found in the recent development of machine learning algorithms involving DL methods, as well as how they compare to prior traditional methods within the field of CV. Currently, DL provides an alternate approach to CV and widens the field of research for CV-related problems. However, traditional CV models prior to DL are still relevant to the field of research. As DL seeks to extract information from an image through the use of automatic feature extraction and selection techniques, the traditional CV approach requires manual involvement in the feature extraction and selection process. Among the different DL models, the Convolutional Neural Network (CNN) is commonly used in the field of CV. Similarly, traditional CV algorithms involving SIFT descriptors show promising results in CV-related tasks. It is of interest in this research effort to highlight where traditional CV algorithms may perform better compared to DL.

R. Y. Gonzalez · S. Latifi (✉)
Department of Electrical and Computer Engineering, University of Nevada Las Vegas, Las Vegas, NV, USA
e-mail: gonzar14@unlv.nevada.edu; shahram.latifi@unlv.edu

According to Wu et al. [2], CV algorithms can be broken down into the following four processes: image pre-processing, feature extraction, feature selection, and prediction/recognition. In traditional CV algorithms, the first three processes have to be manually designed [2]. In the pre-processing stage, an image in the training set may be transformed by means of rotations, shifts, or scaling, in order to easily interpret the relationships and patterns within the image [4]. During the feature extraction stage, features are then extracted from an image in the form of patches of information. This information may include those detected by CV algorithms such as edge detection, corner detection, or threshold segmentation [4]. When features are extracted to form a definition (sometimes referred to as a bag-of-words) for an object class, they are later searched for in other images to determine whether the image is classified as containing that object [4]. However, the main difficulty in traditional CV is that as the number of classes increases, the feature extraction step can become much more complicated to design [4].

In general, feature extraction can be rather involved in CV tasks, such as image classification [4]. Thus, it is up to the designer's judgment to determine which image features are worthy to describe the different classes of objects needed for an application [4]. Since traditional methods separate the feature extraction process from the classifier, it is then important for the designer to select an appropriate classifier algorithm to achieve results best suited for a particular CV task [2].

Unlike the traditional CV approach to manually define and extract a set of features within an image, DL introduces the concept of training from a given image dataset to discover the underlying patterns within different classes of objects [4]. By learning features within the neural networks of a deep learning algorithm, a CV engineer no longer needs to be concerned about the design of the feature extraction process, as they were previously in the traditional approach.

Thus, the two different ML approaches to CV involving extracted hand-crafted features as presented in traditional CV, and automatic learned features as presented in DL, can be summarized in Fig. 61.1 as taken from O'Mahony et al. [4]. It is apparent that from a traditional CV standpoint, there is more control in manually extracting hand-crafted features from images, whereas in DL, the CV engineer must carefully select a training set that can be used to effectively approach a solution to the same problem.

The rest of this paper will be organized beginning with Sect. 61.2, discussing the current state-of-the-art for CV algorithms using automatic feature learning and manual hand-crafted feature extraction/selection techniques. Section 61.3 presents the experimental setups and results obtained throughout three different case studies for the CNN and SIFT

based algorithmic models. A discussion is then made at the end of the three case studies to summarize and extract value from the results obtained in the experiments. Section 61.4 then concludes the report.

61.2 Research on State-of-the-Art for Traditional CV vs. DL

Currently, there are various CV applications that can be solved with the use of different DL algorithms or traditional CV algorithms. Depending on the CV application, one method may be more suitable than another. It is of interest to investigate whether traditional ML algorithms in the field of CV are becoming more obsolete due to advances in DL, or if there is something meaningful that can be extracted from state-of-the-art traditional CV techniques.

61.2.1 Literature Review

According to O'Mahony et al. [4], traditional ML algorithms, using hand-crafted features, can still be considered advantageous compared to DL, depending on the CV application in which certain tasks may be completed in a few lines of programmable code rather than a DL algorithm [4]. Chandrasekhar et al. [5], explains that it has not always been the case that Convolutional Neural Networks (CNNs) have outperformed in CV tasks, such as image retrieval, where hand-crafted features using Scale-Invariant Feature Transform (SIFT) descriptors were found to be comparable to CNNs.

Therefore, a more in-depth analysis will be made to highlight these advantages, from traditional ML algorithms. Before this analysis, it is important to begin with understanding the current state-of-the-art for DL algorithms and traditional ML algorithms, within the field of CV.

61.2.2 Convolutional Neural Network

Presently, there are a variety of DL architectures used within the CV field, which include Generative Adversarial Networks (GANs), CNNs, Recurrent Neural Networks (RNNs), and Fully Convolutional Networks (FCNs) [2, 6]. Of the DL architectures listed, CNNs have become increasingly popular for image processing in CV [2].

The network of a CNN can be broken down into three main layer types: convolutional layers, pooling layers, and fully-connected layers [6]. CNNs adjust weights/parameters inside the layers through a backward propagation approach [2, 6]. In the convolutional layers, an input image is convolved with

Fig. 61.1 (a) Traditional CV and (b) Deep Learning Workflow. (Image by O'Mahony et al. [4])

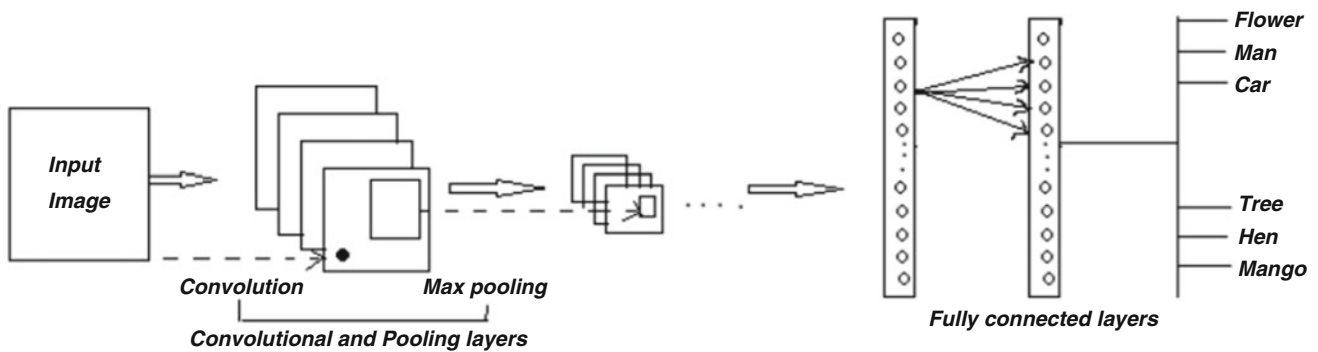
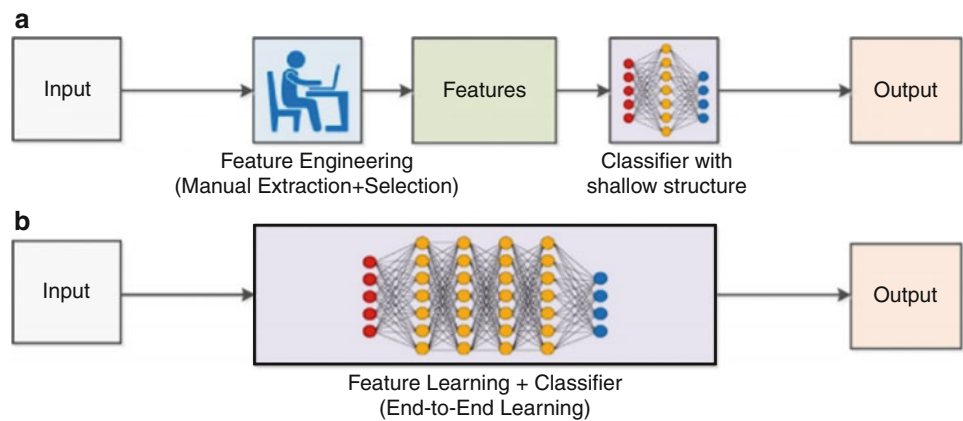


Fig. 61.2 Convolutional neural network pipeline. (Image by R. K. Sinha et al. [6])

filters in order to: reduce the number of parameters in an image using weight sharing mechanisms; find correlation between neighboring pixels; and fix the location of objects within the image [6]. The pooling layers of a CNN will help minimize the measurements of the feature map and parameters within the network [6]. Different pooling approaches include stochastic pooling, spatial pyramid pooling, and def pooling [6]. The fully-connected layers consist of 90% of the parameters as the feed-forward network forms a vector [6]. The architecture pipeline for a CNN is summarized in Fig. 61.2. Currently, different CNN examples include that of the *AlexNet*, *OxfordNet*, and *DenseNet* [3, 5].

61.2.3 Feature Descriptors

Alternatively, in traditional CV algorithms, well-established local feature descriptors such as SIFT, SURF, and BRIEF, have been used for object detection applications [4]. Similarly, global feature descriptors, such as the Fisher Vector (FV), have performed well in image-retrieval CV applications [5, 6]. In general, global and local features extracted by their associated descriptors provide different types of information for an object. Global features may describe an

entire image as a single vector, and thus, provide information that is useful for class discrimination [7]. However, this makes global features sensitive to clutter [7]. Local features may represent information in the form of image patches computed at multiple points within an image [7]. Thus, local features are more robust to occlusion and clutter as compared to global features [7]. In the traditional CV approach, a ML classification algorithm is needed to combine with the feature extraction process as extracted by local and global feature descriptors. The extracted features may be classified by a classifier such as a K-Nearest Neighbor (KNN) or a Support Vector Machine (SVM) algorithm [4]. The KNN classification algorithm has been found in the application of facial recognition in CV [3]. KNN is an easy to learn algorithm that does not require the retraining of newly added data, but suffers from making predictions, due to time complexities [3]. The SVM provides another example of a classifier found in the application of object recognition and image classification in CV [3]. The SVM produces an accurate classifier, despite non-linearities in input data, but does suffer from difficult implementation in large scale datasets [3]. The SVM is also sensitive to missing data, kernel functions selections, and time-consuming hyper parameters [3].

61.2.4 Scale-Invariant Feature Transform

SIFT is a local descriptor that has been shown to be efficient in traditional object-recognition CV applications [8]. SIFT is a feature extraction algorithm used to detect and describe local features in an image that can be used to match with other images [8, 9]. Image matching in traditional CV can be found in applications such as image retrieval [9].

Feature extraction in the SIFT algorithm can be broken down into four steps, which include: estimating a scale space extrema using a Difference of Gaussian; localizing key points/eliminating low contrast points; assigning key point orientations; and computing the local image descriptor for each key point [8]. Feature vectors collected from the SIFT algorithm are aimed at being distinctive and invariant to scaling, rotation, and translation within an image [8]. Furthermore, the first three steps can be categorized as key point detection, and the last step as key point description [9].

As compared to Figs. 61.1a and 61.3 highlights the workflow of a CV algorithm involving the SIFT local descriptor in order to perform image matching [9]. As previously mentioned, after the SIFT algorithm, a classification algorithm, such as KNN or SVM, is then used to perform image matching, based on features extracted from the data [9]. Looking at

Fig. 61.3, one can see that feature filtering is also portrayed in the model. This is typically included to reduce the number of false positives from the data by eliminating irrelevant features according to the application criteria [9].

61.2.5 Fisher Vector

FV is a global descriptor that is also found common in the application of image-retrieval problems within CV [5]. The FV process can be summarized by Chandrasekhar et al. [5]:

The FV is obtained by quantizing the set of local feature descriptors with a small codebook of 64–512 centroids, and aggregating first and second order residual statistics for features quantized to each centroid. The residual statistics from each centroid are concatenated together to obtain the high-dimensional global descriptor representation, typically 8192 to 65,536 dimensions. . . . FVs can be aggregated on descriptors extracted densely in the image, or around interest points like Difference-of-Gaussian (DoG) interest points. The former is popular for image classification tasks, while the latter is used in image retrieval as the DoG interest points provide invariance to scale and rotation [5].

The workflow for the FV process used in a traditional CV algorithm is shown in Fig. 61.4.

Fig. 61.3 Implementing SIFT in a traditional CV model. (Image by Konlambigue et al. [9])

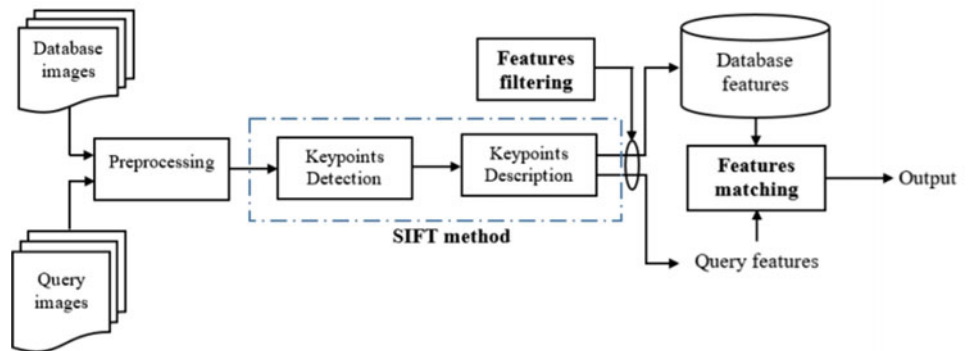
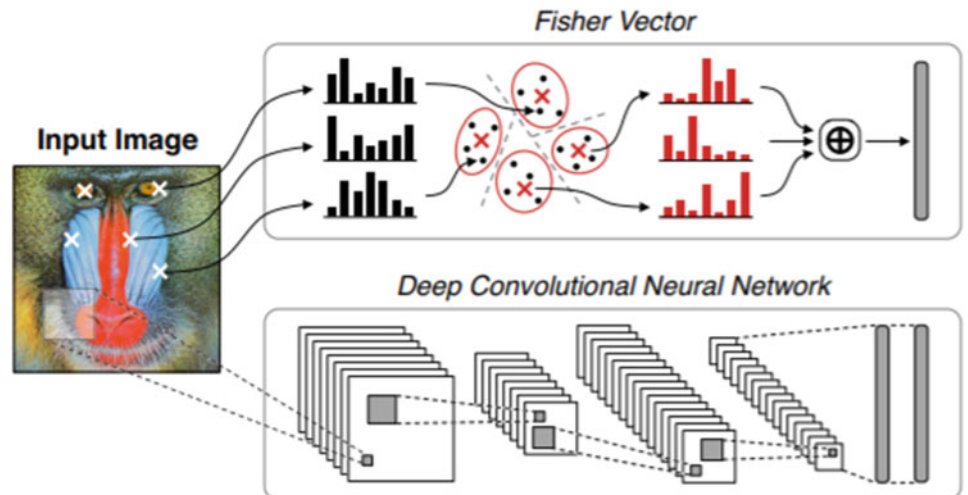


Fig. 61.4 Pipeline of the FV in a traditional CV model. (Image by Chandrasekhar et al. [5])



To summarize the current state-of-the-art in ML CV algorithms, CNNs have been a topic of interest within the field of DL and have found applications in object-recognition. Similarly, traditional CV techniques, involving hand-crafted features extracted from descriptors, are still relevant within the research, more particularly that involving SIFT and FV.

61.3 Case Studies

Now that a general framework has been established to understand the underlying mechanisms involved in the two different ML approaches in CV, an analysis of case studies involving the performance of CNN, SIFT, and FV will be conducted to highlight the pros and cons of the current state-of-the-art. Since CV covers a multitude of different applications, it will be interesting to determine if there are certain applications where traditional CV, such as SIFT and FV, outperforms methods involving deep learning such as the CNNs.

61.3.1 Case Study 1: “Deep Learning Approach for Image Classification”

In [10], a deep learning algorithm involving a CNN is used to binary classify images of cats and dogs from a given image dataset. In this study, a dataset of 10,000 images was collected from a community database known as Kaggle. Among the images, 8000 are used as training data and 2000 for testing. Furthermore, the CNN is trained to learn features from the dataset and is then passed onto a binary classifier. A stochastic gradient descent is used with a binary-cross-entropy cost function to train the network parameters [10]. In the experiment, the CNN is first implemented with two convolutional layers and two pooling layers, with a learning rate set at $\eta = 0.001$.

In the experiment, for 25 epochs, the model gives an accuracy of 84.43% within the training set and 66.1% within the testing set [10]. With 50 epochs, the model gives an accuracy of 84.34% on the training set and 71.35% on the testing set. By adding two more convolutional layers and two pooling layers in the CNN model, for 50 epochs, the accuracy increases to 88.31% for the training set and 84.45% for the testing set [10]. The results are summarized in Fig. 61.5 as taken from the experiment by Panigrahi et al. [10].

Within the study, it indicated that overfitting in the data becomes diminished by adding the two additional convolutional and pooling layers [10]. The authors then conclude that additional training data, along with adding more layers within the CNN could help to improve the test accuracy of the model [10].

Dog or Cat	Precision	Recall	F1-score	support
Class 0 (Cats)	0.79	.77	.77	1000
Class 1 (Dogs)	0.77	0.80	.78	1000
Avg/total	0.78	.78	.77	2000

Dog or Cat	Precision	Recall	F1-score	support
Class 0 (Cats)	0.88	.87	.88	1000
Class 1 (Dogs)	0.87	0.86	.86	1000
Avg/total	0.87	.86	.87	2000

Fig. 61.5 Classification for (a) Top: 25 Epochs, (b) Bottom: 50 Epochs by Panigrahi, et al. [10]

61.3.2 Case Study 2: “Performance Evaluation of State-of-the-Art Filtering Criteria Applied to SIFT Features”

In order for a CNN to perform well, a large amount of data and time is required to train the network [9]. Alternatively, one may use transfer learning in which a pre-trained network is weight adjusted to address a specific task, but this ultimately results in a black box framework [9].

The local descriptor SIFT is among one of the most investigated traditional CV algorithms in object-recognition, image-retrieval, as well as tasks that involve feature matching in general [9]. For the experiment performed by Konlambigue et al. [9], an image retrieval system is implemented using the SIFT algorithm summarized by the CV model in Fig. 61.3 [9]. Different filtering criteria are evaluated on an Oxford-5 k dataset, which contains 5062 high resolution images, distributed in 11 different landmarks, indexed by 55 query images, such that 5 images per landmark are associated in a region of interest (ROI) [9]. The different filtering criteria in the experiment include methods that are Contrast-based (CP), inner primary ratio or IPR-based, Entropy-based, Saliency-based, and descriptor median value filter based [9].

Furthermore, the image matching is done by using a multiple matching removal (MMR) algorithm, where image retrieval performance is then evaluated using a mean average precision (mAP) [9]. The different filtering criteria are then compared by evaluating the change in mAP, relative to the baseline mAP and a CNN mAP [9].

From the results, using a base line SIFT algorithm, the mAP on the Oxford-5 k dataset is approximately 65.91% [9]. Based on the implemented filtering criteria, the experiment

Methods	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	mAP	ρ
SIFT+BoVW[1M]+SP [21]	-	-	-	-	-	-	-	-	-	-	-	64.50	-
SIFT+BoVW[500k]+td-idf [32]	-	-	-	-	-	-	-	-	-	-	-	61.30	-
SIFT+BoVW[20k]+X [33]	-	-	-	-	-	-	-	-	-	-	-	68.50	-
SIFT+BoVW[4096] [4]	36.56	34.25	36.35	41.43	38.17	40.51	69.80	56.03	25.15	56.58	60.24	45.01	-
SIFT [4]	44.90	51.87	51.12	67.24	65.88	67.04	69.24	87.44	27.58	99.18	59.11	62.79	-
CNN [4]	40.36	40.74	39.64	36.18	47.51	45.30	80.65	58.35	27.90	76.45	88.13	52.84	-
Our SIFT baseline (NNDR)	58.64	53.75	50.75	66.59	66.42	61.31	73.18	100	19.75	100	74.68	<u>65.91</u>	0
SIFT+CP[N=5000]	56.33	53.33	50.44	68.42	66.73	61.56	73.21	99.37	19.92	100	74.18	65.77	10.44
SIFT+IPR[th=0.25]	55.88	53.14	52.04	67.32	66.39	58.40	71.95	100	18.74	100	71.42	65.03	10.57
SIFT+Entropy H1[th=5.4]	57.11	51.28	50.43	70.04	66	58.13	69.92	99.37	19.13	100	71.33	64.79	10.07
SIFT+Entropy H2[th=3.5]	58.53	52.60	49.54	69.42	65.70	57.93	69.72	98.04	18.63	100	71.18	64.66	11.45
SIFT+Saliency[th=0.15]	55.79	52.92	49.94	68.92	64.04	62.27	73.69	96.26	19.40	100	70.67	64.90	11.75
SIFT+Median[th=0.007]	58.84	53.33	48.61	68.58	65.61	58.70	70.24	97.84	18.59	100	71.20	64.69	10.52

Fig. 61.6 Summary of results for image retrieval performed by Konlambigue et al. [9]

shows that none of the filters improved the mAP, staying less than 15% below baseline mAP [9]. For reduction rates less than 15%, the contrast-based filtering method is shown to be the best, compared to the other filters [9]. At higher reduction rates, the entropy-based and descriptor median are shown to have the lowest drop in mAP [9]. Reduction rates greater than 40% show that the saliency-based filter was the worst compared to the other filters [9].

Part of the experiment also includes using a CNN approach as opposed to using SIFT based approaches. By using a CNN, the mAP is shown to be approximately 52.84%, which is lower compared to the baseline SIFT and filtered SIFT algorithms [9]. A summary of the results is listed in Fig. 61.6, as taken from the experiment by Konlambigue et al. [9].

61.3.3 Case Study 3: “A Practical Guide to CNNs and Fisher Vectors for Image Instance Retrieval”

In the experimental analysis performed by Chandrasekhar et al. [5], four different datasets are evaluated using a global descriptor FV and a CNN DL algorithm to perform image instance retrieval [5]. The image datasets in the experiment include *INRIA Holidays*, *Oxford Buildings*, *UKBench*, and *Graphics* [5]. The *INRIA Holidays* set includes 500 queries and 991 database images, where variations are rare due to images being taken at the same time [5]. *Oxford Buildings* contains 5062 images of 11 different landmarks, each represented by five possible queries [5]. *UKBench* contains 10,200 images of four different objects with 2550 images, each without foreground and background clutter [5]. *Graphics* contains 1000 database images for 500 unique objects and

1500 queries, with foreground and background clutter to mimic real-world type settings [5].

In the experiment, the datasets are evaluated using an FV with different SIFT descriptors, which include Difference of Gaussian (DoG) SIFT, Dense Single-scale (DS) SIFT, and Dense Multi-scale (DM) SIFT [5]. Similarly, the datasets are also evaluated using different pre-trained CNN models including *OxfordNet*, *AlexNet*, *PlacesNet*, and *HybridNet* [5].

A dimensionality reduction is applied to the SIFT descriptors using a Principal Component Analysis (PCA), and 256 centroids are trained using a Gaussian Mixture Model (GMM) for the FV [5]. The finalized FV is then normalized using power normalization to each component, followed by L_2 normalization [5]. Finally, for the different CNN models, a center cropping strategy is used to achieve results better than padding and squishing crop techniques in an image [5].

The results of the experiment performed by Chandrasekhar et al. [5] are represented in Fig. 61.7, using mean average precision values to evaluate the performance of the FV and CNN models. It is important to note that the *UKBench* dataset contains four different objects in the set, and hence, has a multiplier of four in the mAP, as indicated in the results [5]. From the results, it can be seen that the *INRIA Holidays* and *UKBench* datasets perform better overall with use of the CNN [5]. However, results are better in the *Oxford Buildings* dataset using the FV with DS-SIFT and DM-SIFT descriptors, and are also comparable with the DoG-SIFT descriptors [5]. Furthermore, the FV using the DoG-SIFT descriptors in the *Graphics* dataset reaches mAP at approximately 66%, or roughly a gap of 30% above the different CNN algorithms [5]. Since the *Graphics* dataset contains images of objects at different scales and rotations, the FV containing DoG SIFT descriptors is shown to be more rotation invariant compared to CNN algorithms [5].

Fig. 61.7 Summary of results for image retrieval by Vijay Chandrasekhar et al. [5]

Descriptor	Dim	Holidays	UKBench	Oxbuild	Graphics
<i>OxfordNet</i>	4096	0.80	3.54	0.46	0.33
<i>AlexNet</i>	4096	0.76	3.38	0.42	0.37
<i>HybridNet</i>	4096	0.81	3.39	0.48	0.36
<i>PlacesNet</i>	4096	0.80	3.11	0.46	0.33
CNN (Fine-tuned on Landmarks) [15]	4096	0.793	3.29	0.545	
CNN (Fine-tuned on Objects) [15]	4096	0.754	3.56	0.393	
FVDoG	32768	0.63	2.8	0.42	0.66
FVDS	32768	0.73	2.38	0.51	0.20
FVDM	32768	0.75	2.45	0.55	0.32

61.3.4 Discussion

Based on the first case study for binary image classification system, when training a CNN, one must carefully select an image dataset for a CV system. If one is looking to increase the number of images within the dataset of a CNN, then one should consider the availability of that data for a particular application. As performed in the experiment, an increase in layers to the CNN indicated improvement to the system. However, depending on the requirements of the application, if one seeks to obtain higher accuracy, additional layers may drive complexity to the system, and thus, result in larger computation time compared to a CNN with fewer layers.

In the image retrieval problem performed in the second case study, traditional CV methods involving SIFT algorithms showed better performance compared to the CNN. Additionally, if one plans to use a filtering criterion to the SIFT algorithm, one should take into consideration the affected computation time due to filtering, as well as how it may affect the system design complexity. Furthermore, if one plans to use a feature filter in a SIFT based algorithm, one should also consider the decrease in performance compared to a baseline SIFT algorithm.

Similarly, in the image instance retrieval problem in the third case study as shown by Chandrasekhar et al., it is not always the case that CNNs perform better than traditional CV. The lack of transformation invariance from CNNs poses a drawback compared to FV SIFT descriptor-based algorithms [5]. For the image retrieval problem, if the dataset contains images of different scales and rotations, one may consider using a FV SIFT-based algorithm for better performance. However, the complexity of using a FV SIFT-based design may also play a role in selecting an algorithm for an application, in which pre-trained CNNs that might be able to solve the problem are readily available.

61.4 Conclusions

Recall that traditional CV algorithms, depending on the CV task, may be able to solve a problem more effectively, and possibly, with fewer lines of programmable code. Furthermore, features from traditional CV models require manual involvement in the feature extraction/selection process, as where features learned from deep neural networks require careful selection of a training dataset that is able to extract the features necessary to classify differences in images. It would be unnecessary to utilize a DL algorithm that may be more easily solved using non-class specific CV techniques such as SIFT, color thresholding, and/or pixel counting. Moreover, DL algorithms require very large datasets and high computing power, which are not necessary for traditional CV algorithms.

According to O'Mahoney [4], some CV-related tasks not suited for Deep Learning include those of robotics, augmented reality, automatic panorama stitching, virtual reality, 3D modeling, motion stamation, video stabilization, motion capture, and video processing. Currently, DL performs better for closed-end classification problems; however, it must be supplemented with other techniques to reach artificial intelligence [4]. One drawback in DL is its limited ability to determine whether multiple objects in an image are the same or different, where feedback mechanisms, including perceptual grouping, may be key to realizing visual reasoning [4].

As the growth in technology continues to show advancement in the CV field, DL algorithms are becoming increasingly popular. However, certain applications still rely on traditional CV algorithms, as previously discussed. Three different case studies were highlighted to demonstrate the performance differences for deep learning algorithms using CNNs, and traditional CV algorithms involving the use of

SIFT descriptors. From the case studies, it has been demonstrated that in the image retrieval task, traditional CV methods have the ability to perform better than those resulting from CNNs. Thus, this may drive research to explore how to optimize SIFT algorithms in image retrieval problems, or how to overcome the difficulties faced by CNNs in the same problems.

Acknowledgments A thank you is given to Meagan Madariaga-Hopkins for contribution to proofreading this document.

References

1. R. Szeliski, *Computer Vision Algorithms and Applications*, 1st edn. (Springer-Verlag London Limited, London, 2011), pp. 3–9
2. Q. Wu, Y. Liu, Q. Li, S. Jin, F. Li, The application of deep learning in computer vision. In 2017 Chinese Automation Congress (CAC), pp. 6522–6527 (2017)
3. Z. Weng, From conventional machine learning to AutoML. *J. Phys. Conf. Ser.* **1207**, 012015 (2019)
4. N. O'Mahony et al., Deep learning vs. traditional computer vision. *Adv. Intel. Syst. Comput.* (2020)
5. V. Chandrasekhar et al., A practical guide to CNNs and fisher vectors for image instance retrieval. *Signal Process.* **128**, 426–439 (2016)
6. R.K. Sinha et al., Deep learning for computer vision tasks: a review. (2018)
7. D.A. Lusin, et al. Combining local and global image features for object class recognition. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) – Workshops – Volume 03*. IEEE Computer Society
8. E. Karami, et al., Image identification using SIFT algorithm: performance analysis against different image deformations. (2018)
9. S. Konlambigue, et al., Performance evaluation of state-of-the-art filtering criteria applied to sift features. *2019 IEEE International Symposium on Signal Processing and Information Technology (IS-SPIT)*
10. S. Panigrahi, et al., Deep learning approach for image classification. *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*

Part XII

Data Sciences

Big Data Analytics in Social Media: A Triple T (Types, Techniques, and Taxonomy) Study

62

Md. Saifur Rahman and Hassan Reza

Abstract

Society 2.0; with the help of recent advancements in the internet and web 2.0 technology, makes the social media-based platform the most popular source for big data research. Big Data Analytics contributes by adjusting, analyzing, and forecasting insightful recommendations from this huge source of noisy & mostly unstructured “Big Social Data”. We present 10 mostly used big data analytics in the working domain of social media-based platforms. Different popular techniques or algorithms related to each big data analytic are also listed in this study. We show that “Text Analytics” is the most popular big data analytics in social media data analysis. Through this research, we try to explain the 10 Bigs of big data and introduce the “Sunflower Model of Big Data”. We also explain the reason why the social media-based platform is so significant and popular source of big data by analyzing the most recent statistics. This study will be a handful for all other researchers who want to work with big data in social media and in advance; make their work easy for selecting the best big data analytics method suitable for their research work.

Keywords

Big data · Social media · Big data analytics · Social media analytics · Text analytics · Image analytics · Audio analytics · Video analytics · Predictive analytics · Descriptive analytics · Prescriptive analytics · Web analytics

Md. S. Rahman (✉) · H. Reza
 School of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks, ND, USA
 e-mail: mdsaifur.rahman.1@und.edu; hassan.reza@und.edu

62.1 Introduction

‘Big data’ is an area of great importance for research in the application domain of social media, education, healthcare, complex manufacturing industries, aviation, traffic management, oil & gas exploration, telecommunications, retail, banking & insurance, defense & security, and many more [1].

Statistics from May 2019 show, more than 3.4 billion people around the world are using social media platforms [2] which generate overwhelming structured, semi-structured, unstructured data [3–6] in relatively short timescales. The social media-based platform makes information sharing instant, and speedy. The popularity of social media makes this platform as one of the major communication mediums between the public and emergency responders [7]. Due to the popularity and global reach, many Chief Marketing Officers (CMO) spend some time on social media to give feedback to the product consumers. Statistics shows, 40.8% responded to Twitter, followed by 26.2% on Facebook, and 16.5% on LinkedIn [8]. Therefore, this large volume of data is significant to represent society all over the world. Recent large investments to social media-based big-data-driven decision systems, make this platform traditional alternatives for several organizations [9]. Compare to other traditional marketing service tools, this social media-based platform engaged consumers with the organizations in a better cost-efficient way. As a result, this becomes a great concern for many big organizations to decide on different strategies by analyzing, correlating, and data mining from these big social data sources.

The uprising amount of big social data has brought new challenges into the field of big data analytics. Many new big data analytics have been developed over the last few years into the big data industry to analyze the social data with different data formats like unstructured text, image, audio, video, gif, and blog [4, 5, 10, 11]. These big data analytics are

used for monitoring activity on different social networking websites such as Facebook, Twitter, LinkedIn, Instagram, YouTube, and blogs [1] to obtain profitable insights from this big social data. The selection of the most efficient techniques/algorithms for analyzing this growing amount of big social data is of the utmost market demand.

This paper organizes as follows. (i) Section 62.2 presents the purpose of this study. This section explains the research gap of the previous study and points out the significance of this research work. (ii) Section 62.3 – describes the background study of this work. In this section, we combine the concept of 5 Vs with 10 Bigs to elaborately understand big data. We provide strong statistics on why social media is a growing and most popular source for big data. Besides, we talk about big social data, types of social media, and big data analytics in social media throughout this section of this paper. (iii) Section 62.4 is describing the process flow of this study. The methodology followed, the research questions, and the answers to those research questions in this work. (iv) Section 62.5 is all about the result of this study. First, we show the outcomes of our primary study. Then, we map all techniques/algorithms mostly used with their associative Big Data Analytics in a social media-based platform. We also present supported data types for big data analytics. (v) Through Sect. 62.6, we conclude our study and express our future plan with this study.

62.2 Justification for this Study

Due to the uprising popularity of social media, the implementation of big data analytics, machine learning algorithms, and data science becomes more significant in social media data analysis. Adjustability and variety in big data analytics make this field of data science more suitable for structured, semi-structured, and unstructured data analysis. There is a lot of analytics for big data analysis but not each of these can be implemented in social media for data analysis. Unfortunately, there is a lack of study which presents all possible big data analytics with a variety of implementing techniques or algorithms. To address this gap, we have approached to work on finding the best fit for big data analytics and their associated techniques/algorithms for social media data analysis. We also map each big data analytics with its supporting data type (e.g structured, unstructured). Besides, we describe a broader concept of big data by combining 5 Vs with 10 Bigs. Based on these Bigs we introduce a new model of big data concept named “Sunflower Model of Big Data”. Another purpose of this study is to explain why social media should be considered as a good source of big data. We present the most recent statistics to support this fact. Through this study, we are going to be familiar with big social data, society 2.0, and social data analytics.

62.3 Background Study

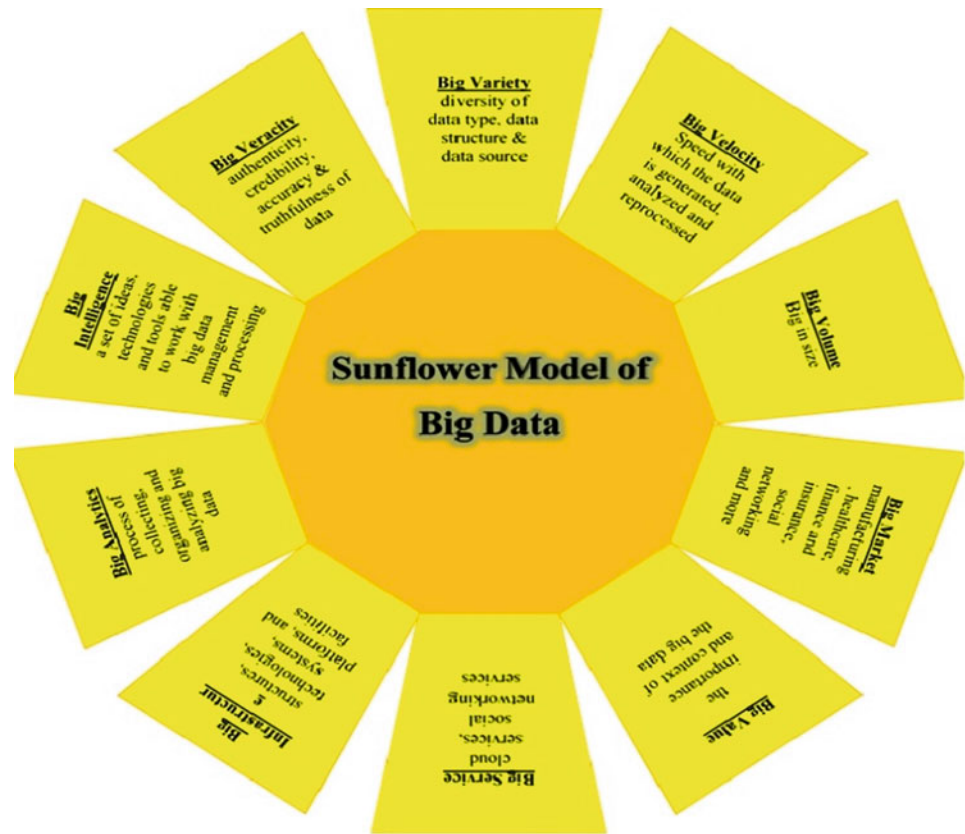
62.3.1 Big Data

Big data defines large and complex data sets including structured, and/or semi-structured and/or mostly unstructured data. Big data requires flexible, and cost-effective data management and analysis techniques to extract insights and make better decisions [12]. Technological advances produce various types of structured but mostly semi-structured or unstructured data daily. Spreadsheets or relational databases produce most of the structured data which is only 5% of existing big data. Online text, images, audio, animation, and video are good examples of unstructured data [11]. Extensible Markup Language (XML) is a good example of semi-structured data [3, 4]. It is very challenging for typical technology to collect, store, manage, and analyze big data because this is too large in volume, too fast, too varied in structure, and too complex. This nature of big data with associated issues challenges current data science algorithms and analytics to resolves the issues [13, 14].

Laney’s (2001) definition of big data based on “3 Vs” (*Volume*, *Velocity*, and *Variety*) is the most accepted and popular definition till now. Later this concept was expanded with 2 more V’s: *Veracity* and *Value*; which are added up by Beyer and Laney (2012) [15, 36]. Most recently, another research group defines big data based on “Bigs” instead of “Vs” [14]. They identify 10 big characteristics related and necessary for big data. The first 4 Bigs (big volume, big velocity, big variety, and big veracity) are fundamental characteristics of big data. The middle 3 Bigs (big intelligence, big analytics, big infrastructure) are technical characteristics of big data. The last 3 Bigs (big service, big value, and big market) are socioeconomic characteristics of big data. We introduce a new model named “Sunflower Model of Big Data” to represent the combination of 5 Vs and 10 Bigs in big data. Figure 62.1 graphically shows the Sunflower Model. A brief description of all Bigs used in the sunflower model is written in Table 62.1.

It is predicted that by 2020 the volume of big data will reach 44 trillion gigabytes [17]. This is because data sharing through social media-based platforms increases every day. Billions of social media users post status updates and share photos, videos on their network daily. These data represent their likes, dislikes, belief, ideas, trends, demographics, and many more [17]. Besides, data generated from a variety of sources like sensor devices, machine logs, bank transactions, mobile communications, geospatial data, and user-generated content in the digital economy should be analyzed [9] to make these unusable data meaningful to someone. Big data analytics helps the researcher in analyzing these

Fig. 62.1 Sunflower model of Big Data



complex and noisy unusable data to extract insightful information for different organizations to make a significant decision.

62.3.2 Social Media

Social Media based platform for multi communication started around 1970 by introducing the PLATO system developed at the University of Illinois [18]. Control Data Corporation first commercially launched social media for public users [18]. Later, a revolutionary change has seen in social media by launching Facebook in 2004 [18]. Facebook allows users to share content and applications that can be flexibly altered by them.

The term “social media” and “social media site” is used interchangeably in this study. Ellison (2007) defines social media based on three important criteria. These are, (i) users network; (ii) allowed to connect with other users form this network; (iii) users can view other user’s public activities within this network [4]. Websites or applications which are used as social networking, social bookmarking, forums, microblogging, social curation, and wikis can be considered as different types of social media [19]. Large amounts of unstructured data, termed as big social data are produced ev-

ery day from Facebook, Twitter, Instagram, LinkedIn, blogs, wikis, YouTube, and many more [4].

The concept of “Society 2.0” was invented based on social media, which is a combination of high-speed internet and web 2.0 technology. Society 2.0 refers to the interaction among people of different real-life societies, through which they can create, share and exchange information ranging from structure to unstructured texts, images, audios, videos, etc. throughout the virtual world. Some of the most popular elements of society 2.0 are Facebook, Twitter, LinkedIn, YouTube, Instagram, Google+, Tumblr, Flickr, emails, forums, blogs, and so on [9]. Figure 62.2 represents a snapshot of social media within 60 seconds in February 2017. The idea about this figure is borrowed from [9] and this figure is collected from online reference [20]. This figure shows a comparative analysis among different social media as a source of big data in different formats and different data types within the years 2014, 2015, and 2016.

The social media platform is using for diverse applications. Table 62.2 shows different types of social media with examples. This table shows, Wikipedia is used for knowledge aggregation; Facebook, Instagram, Myspace for social networking; ResearchGate and Google Scholar for research networking; YouTube and Vimeo for multimedia sharing; Tumblr, Twitter for microblogging, and so on.

Table 62.1 10 Bigs including 5 Vs in Big Data

Bigs/Vs in Big Data	Meaning	Remarks
Big volume	This reflects the size of the data set which is typically in hundreds of terabytes (TB) or petabytes (PB) and even in exabytes (EB) or zettabytes (ZB) [14]. The data volume is relative and varies by factors. Today's big data may not be considered as big data tomorrow because of increasing storage capacities and new technologies [3].	1 st V 2 nd V 3 rd V 4 th V fundamental characteristics
Big velocity	This refers to the speed in terms of data generation, data analysis, and data processing in real-time [16]. It is related to the throughput and latency of data, data in, and data out from a network system in real-time [14].	
Big variety	This refers to the diversity of data types, data structures, and data sources. Big data can be of any data type; structured, semi-structured, and unstructured. In reality, 80% of the data in the world today is unstructured [14, 16].	
Big veracity	This refers to the authenticity and credibility of the data [16]. Accuracy and truthfulness of data are very important in the realm of big data. In reality, there exists ambiguity, incompleteness, uncertainty in big data. We must use big data analytics to remove these complexities from big data [14].	
Big intelligence	This is the artificial intelligence that can be a set of techniques, and tools able to work with big data processing. This is a part of big computing and often works with big intelligence.	technical characteristics
Big analytics	This refers to the process of collecting, organizing, and analyzing big data. The outcome of this process help in decision making based on the discovery of patterns, knowledge extracted from this big volume of data [14]. Big analytics use big intelligence for implementation.	
Big infrastructure	Big data infrastructure refers to the structures, platforms, and facilities which serving big data processing. Apache Hadoop ecosystem can be a good part of big infrastructure [14].	
Big service	This provides services for at least hundreds of millions of people. For example, big data infrastructure services, cloud services, mobile services, big analytics services, social networking services are big services [14].	5 th V socio-economic characteristics
Big value	This indicates the importance and context of big data. The usefulness of big data brings big social value to society in terms of working, living, and thinking [14].	
Big market	The big market includes the market of big data systems, analytics, and services like manufacturing, healthcare, finance and insurance, social networking, and more. This market works on a socio-economical level [14].	



Fig. 62.2 Social Media in 60 seconds [20]

Social Media-based platforms are an unavoidable part of our daily lives which is generating large amounts of data. This big data set can be a good source for data science research. Many business organizations use this big data to make strategic business decisions. Often, users and consumers use social media to search for various information and reviews to make decisions regarding products, education, healthcare, politicians, transportation, insurance, banks, public services, and more. Recently, social media is being used for health status monitoring and mental health applications [26]. “Enterprise 2.0” and businesses are using social media to increase organizational usefulness, improve operational efficiencies, delegate employees, and communication buildup with stakeholders [27]. Indeed business organizations are using different services from Google, YouTube Facebook, Instagram, Twitter, and LinkedIn to stay in business competition by using digital advertisement, raising brand awareness, expanding inbound traffic, and growing search engine optimization [28]. Significantly, many companies are taking advantage of this big social data by applying big data analytics techniques to improve their marketing strategy, customer service activities, and insights extractions [28].

Table 62.2 Types of social media

Types	Social media example
Social networking	Facebook, Myspace [21], Instagram, VKontakte (VK) [22]
Multimedia	YouTube [21], Vimeo, Vine [22]
Knowledge aggregation	Wikipedia [21], Classmates [22]
Microblogging	Twitter [21], Tumblr, Sina Weibo [22]
Instant messaging	WhatsApp, Messenger, Tencent QQ, WeChat, Viber, Line, Snapchat [22]
Professional	LinkedIn, Viadeo, Xing [22]
Forum based	Baidu Tieba (known as Postbar internationally) [22]
Communication	Skype, Hangout, YY [22]
Social bookmarking	Pinterest [22], BibSonomy, CiteULike, Plurk [23], Delicious, Digg, StumbleUpon [24]
Content voting	Reddit [22], Ranker.com, Change.org
Search and discovery	Google, Foursquare [22]
Friendship and dating	Tagged, Badoo [22]
Music-focused	Myspace [22]
Travel and lifestyle	Wayn [22], TripAdvisor [24]
Research	ResearchGate [22], Google Scholar
Blog hosting	WordPress [24]
Opinion and unsolicited reviews	Glassdoor, Zagat, Yelp [24]
Interest-based	Google+ [21], beinggirl.com [24], justmommis.com, alphamom.com, minti.com [25]

62.3.3 Social Media Statistics

Social Media has become a popular platform for research due to a large amount of user-generated content (big data) and the availability of web-based APIs (Application Programming Interface) to access those data. Most social media sites develop their API and provide that to the researcher for working with big social data. For example - Facebook, Google, Instagram, LinkedIn, and many other social media provide their own designed APIs to users but only for research purposes [24].

A popular statistic (June 13th, 2019) from Brandwatch [2] shows, there are 3.499 billion active social media users among 7.7 billion worldwide population. This 45.4% of the total population is almost 80% of active internet users all over the world. This is because social media users grew by 202 million within 1 year (April 2018 to April 2019). Among them, 81% of teenagers feel a positive effect of Social Media on their daily lives. On average, each of these users spent 142 minutes a day on social media. This statistic influences business organizations to invest 74 billion USD on social networking advertisements only in 1 year (2018). That's why 91% of retails brands use social media channels and 81% of

Table 62.3 Statistic of active users in social media

Social media	Active users (monthly)
Facebook	2.375 billion
YouTube	1.9 billion
WhatsApp	1.6 billion
Instagram	1 billion
WeChat	1 billion
LinkedIn	610 million
Weibo	600 million
Twitter	330 million
Pinterest	265 million
Snapchat	190 million
Airbnb	150 million
Google+	111 million
Flickr	90 million
4chan	22 million
Myspace	15 million
ResearchGate	15 million (Wikipedia)
Periscope	10 million

small or medium businesses try to use social media for being on business competition.

Data generated from such a large number of users from social media make this the biggest source of big data. Facebook has become the most popular platform with 2.3 billion monthly active users, YouTube has 1.9 billion active users, WhatsApp has 1.6 billion users, 1 billion users use Instagram, and WeChat and Twitter has 330 million active users in the year 2019. Table 62.3 shows the statistics of active users on different social media platforms [2].

Figure 62.3, shows the percent of active users on different social media platforms. This figure shows that most users on Facebook with 23% of all, then 19% use YouTube, Whatsapp user is 16%, 10% of user use both Instagram & WeChat, and many more.

62.3.4 Big Social Data

Billions of users from different Social Media platforms are continuously generating large amounts of data in the form of text, images, audio, video, gif, blog, and so on [5, 10, 11]. Such a high volume of this structured, semi-structured, and mostly unstructured [19] data should be utilized to extract insightful recommendations and hence make a great decision. This big data set is called "Big Social Data" [27]. Some researchers describe big social data as two types; (i) Social Graph and (ii) Social Text [29]. While another group defines this data set by (i) Interactions (what is being done) and (ii) Conversations (what is being said) [27].

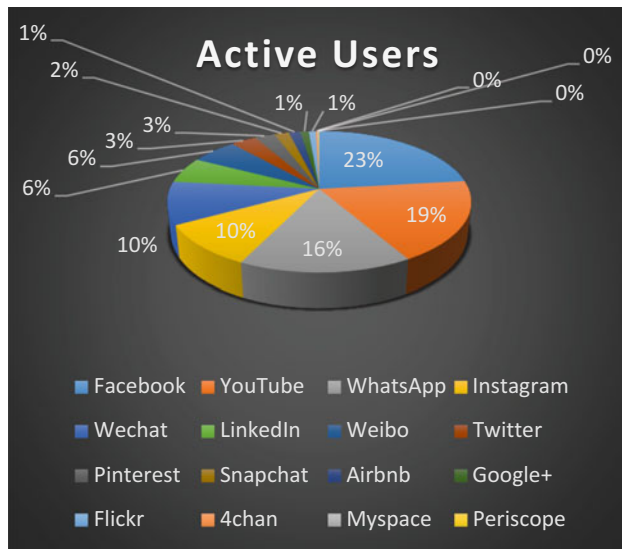


Fig. 62.3 Active users in Social Media in percent

Big Social Data mostly found in the form of unstructured and noisy with complicated social relations; such as follower-followee relationships [21]. Traditional approaches like Relational Database Management Systems (RDBMS) can't analyze this big social data. RDBMS is used to process and analyze a small amount of structured data set [6]. To discover meaningful patterns from big social data, it is compulsory to use any technology which can do computational analysis with the combination of social theories, and statistical & mining methods. The selection of a proper big data analytics tool is the key concern for the social data analysis process.

62.3.5 Big Data Analytics in Social Media

Analytics can be a technique or combination of techniques that can gather, analyze, and extract insightful recommendations from unusable data to make an automatic, faithful, and better decision. Big data analytics can be a sub-process in any knowledge extraction process from big data [3, 11] to discover usable patterns and other recommendations in the decision-making process. These big data analytics are alternatively called "Social Media Analytics" because of their nature of working on various social media platforms only.

Social media analytics work with a dormant understanding of social relationships through interdisciplinary methods like text mining, social network analysis, trend analysis, sentiment analysis, network structure analysis, fraud detection, bioscience, and computational social science applications [21]. To process complex big social data, big data analytics contribute to the development of informatics tools by filtering

and then assembling & extending different suitable machine learning methods [21].

The output of big data analytics can be used by different organizations to improve production or marketing strategy to be alive in digital business competition. Popular business organizations; like Apple, Google, Facebook, eBay, and Amazon regularly use social media analytics for developing their business plan and customer service operations [9]. Governments, healthcare, education, transportation, insurance, academia, and many others come forward to utilize the benefits of social media analytics for strategic planning and making better decisions in an agile manner [9]. These days sensitive events like the local or national election are using social media data analytics to make predictive decisions [5].

62.4 Process Workflow

The flow of this research continues based on the following Research Questions(RQ). RQ1: What are the popular analytics used in the social media-based platform to analyze data? RQ2: Which machine learning algorithms or techniques are used in the implementation of these analytics in social media? RQ3: What are the supporting data types with these algorithms/analytics? RQ4: Which one is the most popular big data analytics in social media data analysis? The primary database for searching relevant research papers was limited to IEEE/IET Electronic Library, ACM Digital Library, and ScienceDirect digital library. By applying several inclusion/exclusion criteria 20 scientific research papers were chosen to continue further study on this topic. The answer to the RQ1 is given in result Sect. 62.5.1 where Table 62.4 presents a list of big data analytics in the social media domain. The result in Sect. 62.5.2 answers the RQ2 by enlisting machine learning algorithms associated with the big data analytics in Table 62.5. Table 62.5 also presents the popularity of that analytics and supporting data types that answer RQ3 and RQ4.

62.5 Results

62.5.1 Popular Big Data Analytics (Found in the Social Media-Based Platform)

This section describes the findings from the primary study of this research work. Table 62.4 organizes and presents the outcomes we get from 20 selected papers used in this study. The left column of this table listed the serial number of the source paper with publication year and reference number. The right column shows the name of the Big Data Analytics (BDA) founded and discussed in that source paper. Analyzing this table, we have found four main categories of BDAs that are frequently used in social media data analysis. These are

Table 62.4 Most frequently used Big Data Analytics in Social Media

Paper ID (publication year)/Reference	Discussed BDAs
S1 (2015), [3]	(i) Text Analytics, (ii) Audio Analytics (Speech Analytics), (iii) Video Analytics (Video Content Analysis – VCA), (iv) Predictive Analytics
S2 (2018), [4]	(i) Descriptive Analytics (Post-mortem Analysis), (ii) Diagnostic Analytics, (iii) Predictive Analytics, (iv) Prescriptive Analytics, (v) Text Analytics
S3 (2014), [1]	(i) Text Analytics, (ii) Web Analytics
S4 (2014), [30]	(i) Text Analytics
S5 (2015), [31]	(i) Video Analytics
S6 (2015), [32]	(i) Text Analytics, (ii) Visual Analytics
S7 (2016), [27]	(i) Visual Analytics, (ii) Predictive Analytics, (iii) Prescriptive Analytics, (iv) Descriptive Analytics, (v) Text Analytics
S8 (2016), [33]	(i) Text Analytics (Social Media Product Improvement Framework (SM-PIF))
S9 (2016), [19]	(i) Text Analytics (Log File Analyzer)
S10 (2016), [34]	(i) Text Analytics, (ii) Visual Analytics
S11 (2018), [6]	(i) Text Analytics
S12 (2019), [7]	(i) Text Analytics, (ii) Image Analytics
S13 (2019), [28]	(i) Text Analytics
S14 (2018), [9]	(i) Text Analytics, (ii) Web Analytics
S15 (2015), [29]	(i) Predictive Analytics, (ii) Descriptive Analytics, (iii) Prescriptive Analytics, (iv) Visual Analytics
S16 (2018), [21]	(i) Text Analytics
S17 (2017), [8]	(i) Text Analytics
S18 (2019), [26]	(i) Text Analytics, (ii) Predictive Analytics
S19 (2016), [35]	(i) Text Analytics
S20 (2018), [24]	(i) Text Analytics, (ii) Image Analytics, (iii) Video Analytics

Text Analytics, Image Analytics, Audio Analytics, and Video Analytics. Besides, we have identified Predictive, Descriptive, Prescriptive, and Diagnostic analytics as independent analytics because of their working behavior. Although, these four are mostly working with text data. At the same time because of the task-specific nature, we include Web Analytics and Visual Analytics in the main category list of big data analytics. So, we have identified the 10 most popular BDAs in the domain of social media for data analysis.

62.5.2 Techniques/Algorithms in Association with BDA in Social Media

Different types of big data analytics in the social media context are related to different techniques or algorithms. This section presents important techniques or machine learning algorithms that are related to big data analytics for structured

and unstructured Big Social Data analysis. Following Table 62.5 listed all those techniques/algorithms which are compatible with and used by 10 BDAs presented in this research work. The first column of the following table listed the serial number and name of the big data analytics with all references who use or discuss those specific analytics in their study. The second column listed techniques/algorithms which are related and can be used with that specific BDA type. The third column shows the popularity of that specific big data analytics in the domain of social media data analysis. The rightmost column shows supported and working data type (structured or unstructured or both) by that technique or BDA type. Among the following techniques, some techniques; like sentiment analysis or artificial neural network, can be used by different BDA types for data analysis. Similarly, some techniques; like Social Set Visualizer (SoSeVi), can be used only by one BDA type (Visual Analytic). These associative techniques play an important role in improving decisions through the analysis of unusable big social data. Thus, these techniques represent a relevant subset of the tools available for big data analytics. We also see that “Text Analytics” is the most popular BDA because 90% of studies from this research refer to these analytics as a computational analysis tool for social data analysis. “Predictive Analytics” is the second most popular BDA which is a variety of “Text Analytics”. “Visual Analytics” is the third popular BDA with 20% popularity. Among these 10 BDAs some work with only unstructured data, where others support both structured and unstructured data types. “Image Analytics”, “Audio Analytics”, “Video Analytics”, “Diagnostic Analytics”, “Prescriptive Analytics”, and “Visual Analytics” mostly work with unstructured data set. On the other hand, “Text Analytics”, “Predictive Analytics”, “Descriptive Analytics”, and “Web Analytics” support both structured and unstructured data types when analyzing social data set.

62.6 Conclusion and Future Work

Many individuals, organizations, and governments use Big Data Analytics to extract valuable insights by analyzing complex big social data. The extracted pattern from this unusable big data set helps to predict or make decisions. This study mostly concentrates on working with big data analytics on the social media platform. We have investigated and listed the 10 most popular big data analytics. We also found that “Text Analytics” is the most frequently used data analysis tool in social media. We explore a list of techniques associate with each big data analytics. Besides, we present supported data types by each of these big data analytics. This research work will surely help social media scientists for their future studies. We expect to continue this research with more scientific study and analysis.

Table 62.5 Different mostly used techniques in association with each Social Media Big Data Analytics

BDA types	Techniques or algorithms	Popularity	Working data type
BDA 1: Text Analytics (Text Mining) / Text Classification [1, 3, 4, 6–9, 19, 21, 24, 26–28, 30, 32–35]	{Information Extraction (Entity Recognition, Relation Extraction)} [3, 27], {Text Summarization (Extractive Approach, Abstractive Approach), Social Influence analysis, link prediction} [3], {Natural Language Processing (Information Retrieval based Approach, Knowledge based Approach, Hybrid Approach)}[3, 24, 35], {Sentiment Analysis/Opinion Mining (Document Level, Sentence Level, Aspect Based, lexical resource approach , Probabilistic Neural Network, Machine-learning Method, Location (Country) Based, Timestamp Based, Followers Count Based, Unstructured Data Normalizer (UDN), Sentiment Analyzer (SA))} [1, 3, 4, 6, 9, 21, 24, 27, 28, 33–35], {SODATO for batch analysis} [34], {Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Gibbs Sampling Approach, Latent Dirichlet Allocation Algorithm, Random Forests (RF), and Decision Tree (DT)}[7], {Support Vector Machine (SVM), Nave Bayesian classifiers (NB)} [7, 28] , {Logistic Regression (LR) algorithms, Multinomial Logistic Regression} [24, 28], {hybrid deep learning structure involving Restricted Boltzmann Machine }[28], {Content Analysis using message, Non-parametric ANOVA Analysis }[8], {Cluster Analysis, Cluster Dendrogram Analysis} [8, 24, 26], {Histogram Analysis, Word Cloud and Commonality analysis, Pyramid Analysis, Cyber Risk Analysis}[26], {Social Network Analysis(social network theory, graph theory)} [3, 4, 21, 24, 27, 32], {Statistical Analysis(Markov chain Monte Carlo methods, regression models, factor analysis)}[24], {Trend Analysis} [21], {Extended Log File Analyzer (cross correlation, self-updating system, customize the configuration, Near Real Time extensions (NRTE))}[19], {Emotional Contagion` Study }[30], {Social Media Product Improvement Framework - SM-PIF (Contextual Information Retrieval (Feature Based Ontology (FBO), Extraction and Storage (ES)), Feature Improvement Recommendation (Product Recommendation Service (PRS))} [33], {Artificial Neural Networks (ANN), Swarm Intelligence, Evolutionary Computation, Deep Learning} [4], {Formal Model, Fuzzy logic} [4, 27, 32].	90%	Structured and Unstructured
BDA 2: Image Analytics (Image Classification) [7, 24]	{Convolutional Neural Networks (CNN), Support Vector Machine (SVM), Linear SVM} [7], {statistical analysis of tag data, demographic data, and download frequency} [24]	10%	Unstructured
BDA 3: Audio Analytics (Speech Analytics) [3]	Transcript-based Approach (large-vocabulary continuous speech recognition, LVCSR) and Phonetic-based Approach [3]	5%	Unstructured
BDA 4: Video Analytics (video content analysis - VCA) [3, 24, 31]	{Server-based Approach and Edge-based Approach} [3], {Modified CCTV VMS (Video Management System) with additional feature of XML format for storage, CCTV metadata analytic} [31], { number of users, response rate, subject, and location}[24]	15%	Unstructured
BDA 5: Predictive Analytics [3, 4, 26, 27, 29]	{Statistical Method} [3], {Modeling, Machine Learning, Game theory} [4], {Google Prediction API}[27], { Social Set Analysis, Linear Regression model based on Social Graph (actors, actions, activities, and artifacts) and Social Text (topics, keywords, pronouns, and sentiments)}[29], {Naive Bayes, K-nearest Neighbors, Support Vector Machines, Decision Trees, Artificial Neural Networks }[26]	25%	Structured and Unstructured
BDA 6: Descriptive analytics (Post-mortem Analysis) [4, 27, 29]	{Work based on historical/past data}[4, 27], {Social Graph Analysis, Social Text Analysis, Social Set Analysis}[29]	15%	Unstructured and Structured
BDA 7: Diagnostic Analytics [4]	Data Discovery, Drill-down, Data Mining, Data Correlations, Data Comprehension, Data Visualization, Search and Filter [4]	5%	Unstructured
BDA 8: Prescriptive Analytics [4, 27, 29]	{Intensive Approach, Optimization, Game Theory, Simulation and Decision techniques}[4], {Social Set Analysis} [27, 29]	15%	Unstructured
BDA 9: Web Analytics [1, 9]	{Google Analytics, AWStats, Amung.us, WebSTAT}[1], {Radian6, Atlas.ti and T-LAB} [9]	10%	Structured and Unstructured
BDA 10: Visual Analytics [27, 29, 32, 34]	{Social Set Visualizer (SoSeVi)} [32], {Social Set Visualizer (SoSeVi) D3.js; a JavaScript based visualization framework} [27], {visual analytics tool Tableau, Social set visualizer (SoSeVi)} [34], {SSA approach using D3.js libraries, Social Data Analytics Tool (SODATO), Social Set Visualizer} [29]	20%	Unstructured

References

- Z. Dhawan, Big data and social media analytics. Res. Matters A Cambridge Assess. Publ. **18**, 36–41 (2014)
- K. Smith, 126 Amazing Social Media Statistics and Facts, [brandwatch.com](https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/), 2019. [Online]. Available: <https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/>. Accessed: 03-Aug-2019
- A. Gandomi, M. Haider, Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manag. **35**(2), 137–144 (2015)
- N.A. Ghani, S. Hamid, I.A. Targio Hashem, E. Ahmed, Social media big data analytics: a survey. Comput. Human Behav. **101**, 417–428 (2018)
- P. Victor Paul, K. Monica, M. Trishanka, A survey on big data analytics using social media data, *2017 Innov. Power Adv. Comput. Technol. i-PACT 2017*, vol. 2017-Janua, pp. 1–4 (2018)
- F. Shaikh, F. Rangrez, A. Khan, U. Shaikh, Social media analytics based on big data, *Proc. 2017 Int. Conf. Intell. Comput. Control. I2C2 2017*, vol. 2018-Janua, pp. 1–6 (2018)
- V. Nunavath, M. Goodwin, The role of artificial intelligence in social media big data analytics for disaster management -initial results of a systematic literature review, *2018 5th Int. Conf. Inf. Commun. Technol. Disaster Manag. ICT-DM 2018*, no. M1, pp. 1–4 (2019)
- F. Piccialli, J.E. Jung, Understanding customer experience diffusion on social networking services by big data analytics. Mob. Networks Appl. **22**(4), 605–612 (2017)
- P. Ducange, R. Pecori, P. Mezzina, A glimpse on big data analytics in the framework of marketing strategies. Soft. Comput. **22**(1), 325–342 (2018)
- P. Grover, A.K. Kar, Big data analytics: a review on theoretical contributions and tools used in literature. Glob. J. Flex. Syst. Manag. **18**(3), 203–229 (2017)
- J. Amudhavel, V. Padmapriya, V. Gowri, K. LakshmiPriya, K.P. Kumar, B. Thiyagarajan, Perspectives, motivations, and implications of big data analytics, pp. 1–5 (2015)
- M. Gupta, J.F. George, Toward the development of a big data analytics capability. Inf. Manag. **53**(8), 1049–1064 (2016)
- L. Cao, Data science: challenges and directions, pp. 59–68 (2017)
- Z. Sun, K. Strang, Big data with ten big characteristics, (2018)
- B. Sena, B. Sena, A.P. Allian, E.Y. Nakagawa, Characterizing big data software architectures: a systematic mapping study, no. September 2017
- What is Big Data? – A definition with five Vs, [blog.unbelievable-machine.com](https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs/), 2018. [Online]. Available: [https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs.](https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs/) [Accessed: 07-Aug-2019]
- N. Dave, 4 major ways in which big data is impacting social media marketing, [insidebigdata.com](https://insidebigdata.com/2018/10/06/4-major-ways-big-data-impacting-social-media-marketing/), 2018. [Online]. Available: <https://insidebigdata.com/2018/10/06/4-major-ways-big-data-impacting-social-media-marketing/>
- W. Contributors, Timeline of social media, [en.wikipedia.org](https://en.wikipedia.org/wiki/Timeline_of_social_media), 2019. [Online]. Available: https://en.wikipedia.org/wiki/Timeline_of_social_media. [Accessed: 07-Aug-2019]
- C.J. Aivalis, K. Gatzliolis, A.C. Boucouvalas, Evolving analytics for e-commerce applications: utilizing big data and social media extensions, *2016 Int. Conf. Telecommun. Multimedia, TEMU 2016*, pp. 188–193 (2016)
- R. Allen, What happens online in 60 seconds? [smartinsights.com](https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/), 2017. [Online]. Available: <https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/>. [Accessed: 07-Oct-2019]
- W.Y. Ayele, G. Juell-Skielse, Social media analytics and internet of things, pp. 1–11 (2018)
- J. Spencer, 65+ Social Networking Sites You Need to Know About, [makeawebsitehub.com](https://makeawebsitehub.com/social-media-sites/), 2019. [Online]. Available: <https://makeawebsitehub.com/social-media-sites/>. [Accessed: 07-Oct-2019]
- W. Contributors, List of social bookmarking websites, [en.wikipedia.org](https://en.wikipedia.org/wiki/List_of_social_bookmarking_websites), 2019. [Online]. Available: https://en.wikipedia.org/wiki/List_of_social_bookmarking_websites. [Accessed: 07-Oct-2019]
- I. Lee, Social media analytics for enterprises: typology, methods, and processes. Bus. Horiz. **61**(2), 199–210 (2018)
- The Best Social Networks for Moms, [ranker.com](https://www.ranker.com/list/mom-social-networks/liz-paddington). [Online]. Available: <https://www.ranker.com/list/mom-social-networks/liz-paddington>. [Accessed: 07-Oct-2019]
- A. Subroto, A. Apriyana, Cyber risk prediction through social media big data analytics and statistical machine learning. J. Big Data **6**(1), 50 (2019)
- R. Vatrappu, R.R. Mukkamala, A. Hussain, B. Flesch, Social set analysis: a set theoretical approach to big data analytics. IEEE Access **4**, 2542–2571 (2016)
- M. Ngaboyamahina, S. Yi, The impact of sentiment analysis on social media to assess customer satisfaction: case of rwanda, *2019 IEEE 4th Int. Conf. Big Data Anal.*, pp. 356–359 (2019)
- R. Vatrappu, A. Hussain, N. B. Lassen, R.R. Mukkamala, B. Flesch, R. Madsen, Social set analysis: four demonstrative case studies, pp. 1–9 (2015)
- R. Schroeder, Big Data and the brave new world of social media research. Big Data Soc. **2**, 1 (2014)
- K. Park, M.C. Nguyen, H. Won, Web-based collaborative big data analytics on big data as a service platform, *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2015–August, pp. 564–567 (2015)
- B. Flesch, R. Vatrappu, R.R. Mukkamala, A. Hussain, Social set visualizer: a set-theoretical approach to big social data analytics of real-world events, *Proc. – 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, pp. 2418–2427 (2015)
- C.J. Su, Y.A. Chen, Social media analytics based product improvement framework, *Proc. – 2016 IEEE Int. Symp. Comput. Consum. Control. IS3C 2016*, pp. 393–396 (2016)
- A. Hennig et al., Big social data analytics of changes in consumer behaviour and opinion of a TV broadcaster, *Proc. – 2016 IEEE Int. Conf. Big Data, Big Data 2016*, pp. 3839–3848 (2016)
- M. Conway, D. O’Connor, Social media, big data, and mental health: current advances and ethical implications. Curr. Opin. Psychol. **9**, 77–82 (2016)
- M.S. Rahman, H. Reza, Systematic mapping study of non-functional requirements in big data system, *2020 IEEE International Conference on Electro Information Technology (EIT)*, Chicago, IL, USA, 2020, pp. 025–031, <https://doi.org/10.1109/EIT48999.2020.9208288>

CARS: A Containerized Amazon Recommender System

63

Adam Cassell, Andrew Muñoz, Brianna Blain-Castelli, Nikkolas Irwin, Feng Yan, Sergiu M. Dascalu, and Frederick C. Harris Jr.

Abstract

With the big data boom, recommender systems that make intelligent recommendations for users have been playing an important role in today's industry. However, existing recommender systems often overlook scalability, flexibility, and portability. They also commonly lack in-situ visualizations. To solve these problems, we present CARS: A Containerized Amazon Recommender System. CARS processes large Amazon data sets for analysis and makes product recommendations. However, its utility is not restricted to only prominent organizations like Amazon. CARS achieves scalability by taking advantage of industry-grade recommendation tools irrespective of available hardware resources. CARS runs in a completely isolated environment to promote flexibility and remote collaboration. The demonstrated implementation generates shopping recommendations from user ratings within product review data sets. CARS processes this review data using Apache Spark, a unified analytics engine for big data. The system complements recommendations with data-driven insights and interactive visualizations. In addition to these features, CARS contains a robust set of command line options to customize the results shown to the end-user, perform logging of processed data, and provide performance monitoring through Spark's built-in web-interface. Highly portable and automated analysis of purchase data helps organizations understand the habits of their customers. CARS demonstrates the feasibility of such a system for a wide variety of users.

A. Cassell · A. Muñoz · B. Blain-Castelli · N. Irwin · F. Yan
S. M. Dascalu · F. C. Harris Jr. (✉)
Computer Science and Engineering, University of Nevada, Reno,
Reno, NV, USA
e-mail: acassell@nevada.unr.edu; amunoz24@nevada.unr.edu;
bblaincastelli@unr.edu; nikkolasjirwin@nevada.unr.edu;
fyan@unr.edu; dascalus@cse.unr.edu; fred.harris@cse.unr.edu

Keywords

Big data · Scalable systems · Containerization · Recommender system · Data analysis · Data visualization · Collaborative filtering · Alternating least squares · Apache spark · Performance analysis

63.1 Introduction

As companies continue to accumulate big data, it is important to use the collected data to enhance customer experiences and help make data-driven decisions. Recommendation systems are commonly found on e-commerce and entertainment websites. These systems collect individual behavioral data including purchases, ratings, and views. This data, once gathered, is then processed and used to provide future recommendations to users with similar interests.

For large organizations like Amazon and Netflix, the collection of user data, processing of user data, and use of robust recommender systems can be easily accomplished due to large development teams and support infrastructure. However, smaller organizations do not have such resources even though they also want to benefit from these systems. Additionally, the dynamic nature of web traffic makes it difficult for organizations to manage data at scale without building applications that are designed for cloud environments.

CARS uses a containerized environment to provide an accessible, light-weight, portable and scalable recommendation system. We prototype CARS on top of Conda, Docker, Python, and Spark. Our extensive evaluation of CARS demonstrates promising results. CARS allows review data to be processed in a wide range of environments using minimal hardware that can be scaled up to meet the needs of the end-user. The two fixed constraints for CARS are the use of Docker and the physical resources

allocated to the container. Item recommendation results and insightful aggregated statistics are provided to users alongside interactive visualizations which provide fine-grained analysis.

The rest of this paper is organized as follows: Sect. 63.2 discusses the project background and provides examples and information on related work. Section 63.3 describes the implementation of CARS. Section 63.4 evaluates the individual visualizations and analysis. Section 63.5 lists concluding thoughts and future work.

63.2 Background and Related Work

According to *Spark the Definitive Guide*, recommender systems are “one of the best use cases for big data” [1]. Such a system can analyze users’ explicit preferences or implicit preferences to make predictions on a user’s future behavior. There are many use cases for recommender systems across a variety of domains. Entertainment platforms often use such systems to help users discover new content that they may like. For example, Netflix leverages Apache Spark to implement the movie recommendation system that many consumers are familiar with. Amazon, likewise, uses a similar implementation on its item catalog to drive shopping recommendations based on user preferences and interests. A previous user study shows the significance of these recommender systems on user shopping trends at Amazon when “[a]ll participants used one or more recommender feature[s]” [2]. User-data-driven recommendations have become core components of many large-scale products. It follows that making these processes accessible to a wider variety of systems will bring more effective experiences to even more domains.

One way to implement a recommender system is by using a content-based approach, which uses detailed information about specific items or user attributes to drive predictions. This includes features such as textual content, genre, item category, and other relevant metadata [3]. Conversely, collaborative filtering relies on the historical preferences of users (i.e. the actions they take). This approach is especially useful for “complex and hard to represent concepts, such as taste and quality” [4].

There are two types of user preferences: explicit rating and implicit rating. The former describes direct ratings given by users, such as a five-star rating or numeric scale. The latter is characterized by indirect indications, such as page clicks, image views, purchase records, and other passively traceable statistics. CARS uses explicit rating for user preferences.

One of the most widely-used algorithms for collaborative filtering is Alternating Least Squares (ALS), which supports

explicit or implicit user feedback. Spark natively supports ALS and includes multiple variants of scalable ALS implementations in its MLlib library. MLlib was used for the CARS recommender system due to its convenience and performance benefits. ALS finds a k -dimensional feature vector for every user and item such that the dot product of each item’s feature vector with each user’s feature vector estimates the user’s rating for that item [1]. This resulting feature vector then drives the recommendations.

A set of Amazon’s review data is made freely available by researchers at UCSD [5]. This data set is divided into separate subsets of data which include the five-core data sets, raw review data, user review data, and ratings only data. As stated previously, the CARS project utilized the five-core data sets for its implementation. This Amazon data set has also been the subject of multiple related works. In the paper *Estimating Reactions And Recommending Products With Generative Models Of Reviews*, Ni et al. use the Amazon data set to generate predicted review text using natural language inference techniques [6].

Another paper, *Justifying Recommendations using Distantly-Labeled Reviews and Finer-grained Aspects* by Ni et al., uses Amazon clothing data to extract meaningful justifications that are pertinent to customers’ decision-making processes [7]. An additional work, titled *Large Scale Parallel Collaborative Filtering for the Netflix Prize* by Zhou et al. [8], uses Netflix data to demonstrate an ALS implementation with weighted regularization for ratings prediction. It is also worth noting that both papers, [6] and [7], were written by the researchers who made the Amazon review data set freely available.

63.3 Approach

The implementation of CARS starts with Conda. First, dependencies are installed and then the application is isolated into a Docker container equipped with Jupyter Notebook. Then, the recommender system and interactive visualizations are added. After completing the features above, a command line parser is integrated to provide the end-user with additional options when executing CARS with a given data set.

CARS uses Conda, “an open source package management system and environment management system that runs on Windows, macOS and Linux” [9], to facilitate cross-platform package management, dependency management, and isolation through a virtual environment. By using Conda, CARS automates the process of installing and running the application in a consistent and reproducible manner. Through Conda, CARS users can also extend the existing application by installing new packages without dependency conflicts.

Conda, and the application code for CARS are contained within a Docker container. This container uses a custom image that builds upon Project Jupyter's minimal-notebook image. Using this approach, CARS was guaranteed to have stable Jupyter Notebook support along with custom configurations. One of the configurations required by CARS was the ability to store results regardless of the container's lifecycle/state.

Relying on the container's persistence layer would not be sufficient for working with the Amazon review data sets since any work performed while running CARS would be lost if the CARS container's lifecycle was interrupted. The solution to this issue was to add a mount point to the custom Docker image file and then configure the volume to store our files as well as the data sets used for running our recommender system.

While adding a Docker volume may not be a necessity for end-users who have experience with Docker, CARS utilizes this mechanism to bundle the data sets directly into CARS so that the end-user can immediately begin working with the data sets and further customizing the program for their specific needs. To ensure that the Docker image size is not too large, only a subset of the five-core data sets is bundled by default, but more can be added to the Docker volume as needed.

The data visualizations were designed to serve as key insight into the Amazon review data and user preferences. The visualizations that were created include:

- Items Over Time
- Summary Statistics
- Helpful Reviews
- Prediction Performance

These data visualizations provide insight into some of the basic statistics of the review data, trends over time, and several interesting relationships. The prediction performance plots explore how well the ALS algorithm handles the data and produces its recommendations. Each of the data visualizations mentioned will be explained in further detail in Sect. 63.4.

63.4 Evaluation

We evaluate CARS in this section, including review data analysis and recommendation performance. Corresponding visualizations serve to highlight different aspects of the data, such as relationships or changes over a period of time. All visualizations depicted and shown throughout this section utilize the five-core Video Games data set as the primary

basis. The five-core data sets is simply a collection of products and reviewers that each have a minimum of at least five reviews. This helped with processing and running the data sets locally.

63.4.1 Items Over Time

It is possible to infer an item's popularity over time using review data. For any data set provided, CARS selects the most popular item of that category as a case study for such analysis. The result is a time-series line plot, which can be seen in Fig. 63.1, displaying the number of daily reviews for that particular item.

It is worthwhile to observe how an item's popularity changes over time as potential indicators of product success. In this example, this particular video game received a large amount of reviews (over 100) on the first day of release. After that, however, popularity rapidly dropped after those initial few days and weeks (down to more modest single-digit daily numbers). This visual suggests that the item had high public interest before and at the time of release. This pattern intuitively makes sense for something like a video game or movie. There is built-up demand followed by the item quickly becoming outdated compared to other new releases. Contrast this with an item of another category, such as medical. Popularity for a medical item, such as a box of bandages, is likely to have much more steady popularity as it is a common good. Marketers could use these insights to best plan their advertising strategies. Advertising campaigns could either take advantage of the existing popularity patterns, or attempt to change them to better reach certain goals.

63.4.2 Summary Statistics

It is important to evaluate the distribution of the data set when considering any downstream analysis. Table 63.1 shows the Amazon data set schema and a subset of the video games data. Summary statistics are provided in the form of a summary table as well as a ratings distribution histogram. The summary table, as seen in Table 63.2, communicates the mean, standard deviation, count, minimum/maximum values, and quartile ranges of all rating values in the data set. The ratings distribution displayed in Fig. 63.2 shows how many occurrences of each rating (discrete values between one and five) exist in the data set.

The histogram is useful for understanding how the ratings are distributed for each data set. In this case, it is evident that video game ratings overwhelmingly trend positive, with the highest occurring ratings being five stars, followed by four star ratings.

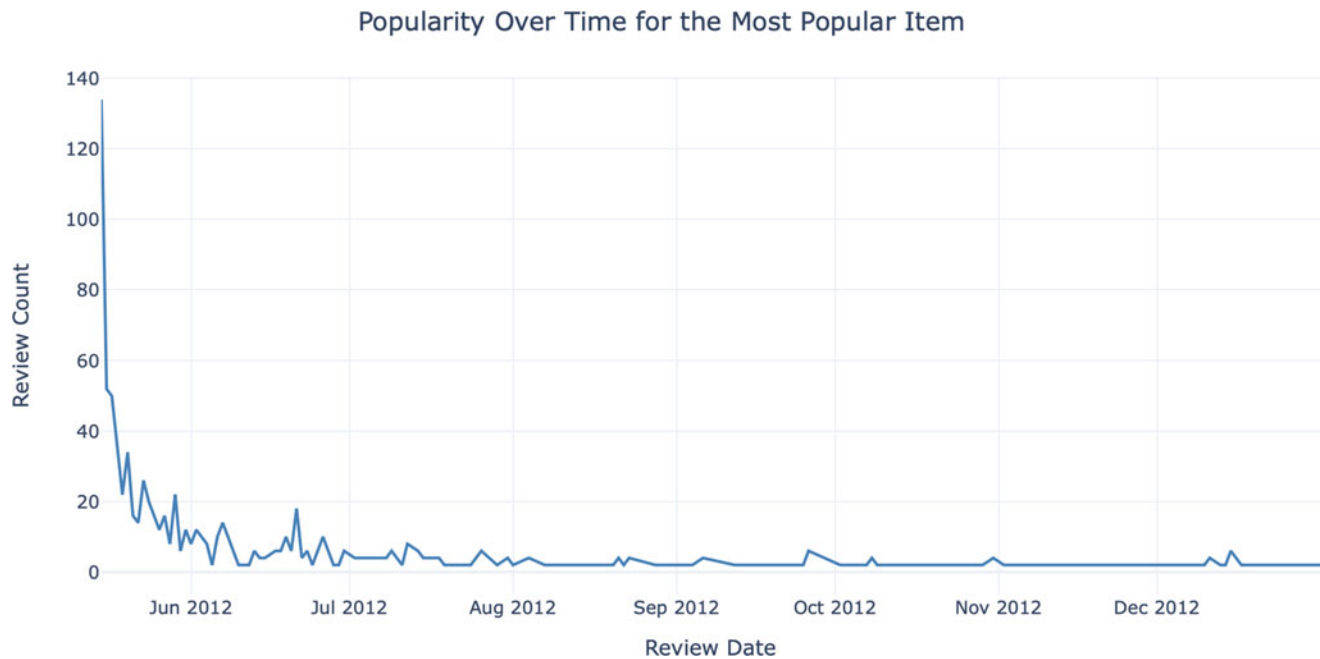


Fig. 63.1 Time series plot showing review count over time for the most popular item

63.4.3 Helpful Review Data

A valuable metric for analyzing user reviews is review ‘up-votes’. Users can express how ‘helpful’ they find a particular review by voting on it. This in turn helps shoppers decide which reviews to give more credence to, thus revealing the most influential reviews. CARS visualizes this metric for the most-reviewed item of the data set, as well as for the data set as a whole. This evaluation section focuses on the former visualization for the most popular item.

To explore the ratings and votes associated with each of the item’s reviews, a scatter plot is provided in Fig. 63.3. The purpose of this plot is mostly investigative. The x -axis represents every single reviewer for the item, and the dual y -axes represent ratings and votes, respectively. The intended usage is for the to pan and zoom around this plot to analyze items of interest. This plot, like many others generated by CARS, includes Plotly [10] interactivity for these purposes. This particular visualization implements WebGL to more efficiently render hundreds of thousands of interactive points if necessitated by the data set.

Using this visual, it is evident that most users up-voted the reviews that gave one-star ratings. This implies a generally negative shopper consensus. Very few shoppers up-voted the five-star reviews, which is notable. There is a stark contrast between this item’s vote distribution and the general ratings distribution for the video game category as a whole (see Sect. 63.4.2). A logical conclusion from this discrepancy

follows: This item seems to be far more poorly received than most of its competitors. Yet, it was still far more popular than the other games sold on Amazon. It is up to the seller to determine whether this constitutes a success.

63.4.4 Prediction Performance

The previous visualizations pertained to general analysis of the data set. The following outputs instead demonstrate the performance of the ALS algorithm used for the recommender system in CARS. Before ALS can generate its list of item recommendations per user, rating predictions must first be computed. These are the rating values that the algorithm predicts each user would assign to each item. These predicted ratings form the basis of the eventual recommendation groupings presented as final output. Thus, the accuracy of these predictions is critical and can be visualized to assess ALS performance.

To best visualize prediction accuracy, Fig. 63.4 shows an error distribution plot. Prediction error is calculated by subtracting the true rating from the predicted rating for each sample. The result is a prediction error sequence that can be organized using a histogram. The resulting distribution is close to Gaussian, which indicates the algorithm is satisfyingly accurate. In Table 63.3, the final recommendation results outputted by CARS are listed.

Table 63.1 Structure and example data of the video game review data set

ASIN	Rating	Review text	Review time	Reviewer ID	Reviewer name	Summary	Verified	Vote
0700026657	5	This game is a bit hard to get the hang but when you do it's great.	10 17, 2015	A1HP7NVNPFMA4N	Ambrosia075	But when you do it's great.	True	Null
0700026657	4	I played it a while but it was alright. The steam was a bit of trouble. The more th move these game to steam the more of hard time I have activating and playing game. But in spite of that it was fun, 11 it. Now I am looking forward to anno 22 really want to play my way to the moon	07 27, 2015	A1JGAP0185YJ16	travis	But in spite of that it was fun, I liked it	False	Null
0700026657	3	ok game.	02 23, 2015	A1YJWEXHQBWK2B	Vincent G. Mezera	Three stars	True	Null
0700026657	2	Found the game a bit too complicated, n what I expected after having played 16C 1503, and 1701	02 20, 2015	A2204E1TH211HT	Grandma KR	Two stars	True	Null
0700026657	5	Great game, I love it and have played it since its arrived	12 25, 2014	A2RF5B5H74JLPE	jon	Love this game	True	Null

Table 63.2 Summary statistics of the video game review data set

Metric	Overall
Count	497577
Mean	4.220456331381876
Std	1.1854244331373522
Min	1
25%	4
50%	5
75%	5
Max	5

Fig. 63.2 Histogram visualizing the ratings distribution of the video game review data set



Ratings and Votes Concentrations for the Most Popular Item

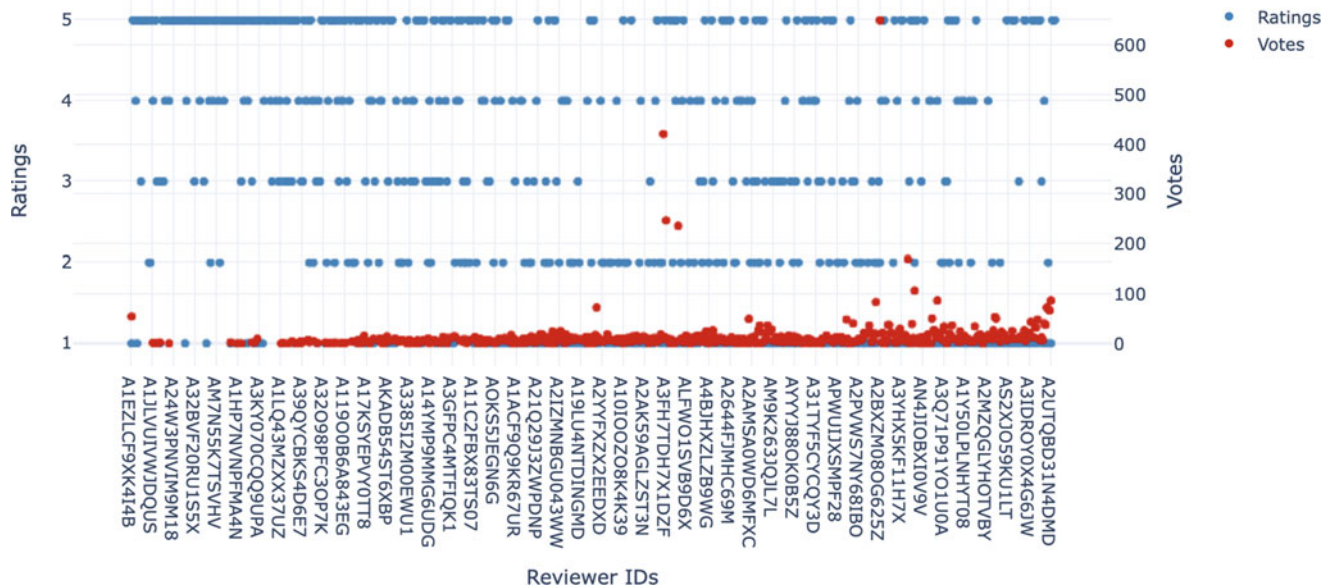


Fig. 63.3 Ratings and votes concentration for the most popular item in the video game review data set

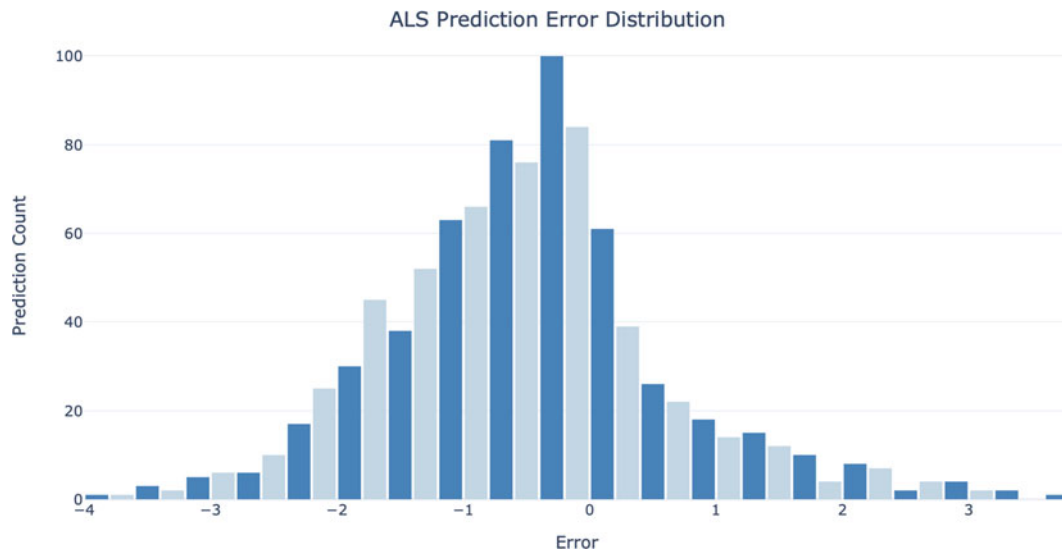


Fig. 63.4 Error distribution plot demonstrating ALS prediction performance in CARS

Table 63.3 Recommendation results computed by ALS algorithm

Reviewer ID	ASIN, rating
148	14149,6.55188512802124
463	5803,6.189663887023926
471	17321,6.773868560791016
496	11845,8.053912162780762
833	16655,7.285752773284912

63.5 Conclusion and Future Work

Recommender systems enhance user experiences by providing suggestions tailored towards users' interests. Implementing recommender systems and collecting enough user data to generate accurate recommendations can be challenging. CARS is designed to address these challenges and demonstrates the feasibility of such a system for organizations of any size. Built on top of Conda, Docker, Python, and Spark, CARS enables extensibility, dependability, portability, and scalability for systems deployed on varying hardware resources. In addition to these features, CARS supports a variety of visualizations on individual product and aggregated data.

As our future work, CARS will be improved by incorporating Ansible into the design so that our containerized application can be further automated to simplify the executions. In addition to automated Docker, CARS will continue to gain new data visualization features. The visualization wish list includes sales rank, item popularity based on categories, and comparative analysis based on different parameters.

Acknowledgments This material is based in part upon work supported by the National Science Foundation under grant numbers IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. B. Chambers, M. Zaharia, *Spark: The Definitive Guide : Big Data Processing Made Simple* (O'Reilly Media, Newton, 2018), pp. 1–11, 468–475
2. J. Leino, K.-J. Räihä, Case amazon: ratings and reviews as part of recommendations, in *Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys '07*, New York, NY (Association for Computing Machinery, New York, 2007), pp. 137–140
3. S. Luo, Introduction to recommender System. <https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26>, December 2018. Last accessed 24 March 2020
4. J.L. Herlocker, J.A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, CSCW '00*, New York, NY (Association for Computing Machinery, New York, 2000), pp. 241–250
5. J. Ni, Amazon Review Data. <https://nijianmo.github.io/amazon/index.html> (2018). Accessed 24 March 2020
6. J. Ni, Z.C. Lipton, S. Vikram, J. McAuley, Estimating reactions and recommending products with generative models of reviews, in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Taipei (Asian Federation of Natural Language Processing, Canberra, 2017), pp. 783–791
7. J. Ni, J. Li, J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

- Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong (Association for Computational Linguistics, Stroudsburg, 2019), pp. 188–197
8. Y. Zhou, D. Wilkinson, R. Schreiber, R. Pan, Large-scale parallel collaborative filtering for the netflix prize, in *Algorithmic Aspects in Information and Management*, ed. by R. Fleischer, J. Xu (Springer, Berlin, Heidelberg, 2008), pp. 337–348
 9. Anaconda, Inc., Conda. <https://docs.conda.io/en/latest/> (2017) Accessed 14 May 2020
 10. Plotly, Take data science and AI out of the lab. Free the data. Share the knowledge. <https://plotly.com> (2020) Last accessed 24 March 2020

Using Technologies to Uncover Patterns in Human Trafficking

Annamaria Szakonyi, Harshini Chellasamy, Andreas Vassilakos, and Maurice Dawson

Abstract

In this paper, the researchers provide a background of human trafficking, review the use and implications of digital currency, and apply machine learning techniques to analyze publicly available trafficking datasets. The study also provides recommendations related to data collection, management, and analysis to aid the vital fight against individuals' exploitation. The researchers conducted an exploratory data analysis using Python, RapidMiner, and Microsoft Excel towards an iterative review and interpretation of the dataset from the Counter Trafficking Data Collaborative (CTDC). The researchers found that there are more female victims of human trafficking in most age groups than male victims. However, for the age group between 39–47, there was a higher male victim count. Additionally, researchers found that the top five countries affected by human trafficking were the Philippines, Ukraine, Republic of Moldova, USA, and Cambodia. However, it must be noted that there are limitations to the overall data because they are provided voluntarily by organizations, and therefore, there is no equitable distribution of actual results from all countries and players. After mapping the country of origin and country of exploitation, it was made clear that there is a movement of victims from the country of origin to the country of exploitation. Lastly, researchers found that a complex combination of different variables is needed to provide accurate predictions for law enforcement and anti-trafficking organizations to aid them in fighting human trafficking, including country

of exploitation and type of exploitation being the most important features in the prediction.

Keywords

Human trafficking · Law enforcement · Dark web · Metadata · Blockchain · Bitcoin · Cryptocurrency · Data analysis · Data science · Machine learning · Technology · Gender

64.1 Introduction

Human trafficking, also known as modern-day slavery, is the use of force or coercion to exploit children and adults for labor or sex [1]. In 2016, modern-day slavery affected roughly 40.3 million people around the world [2]. It is essential that human trafficking is tracked down because the crime results in catastrophic consequences for its victims and the societies that it infects. Victims of human trafficking often experience numerous physical and psychological problems including disease, clinical depression, and anxiety disorders [3, 4]. As a result, human trafficking is becoming increasingly recognized as a public health problem by many organizations including the American Academy of Family Physicians and the American Psychological Association [5].

Human trafficking is a national security threat to the countries that it plagues as it works in ways similar to drugs and arms trafficking [6]. Human traffickers have a deep understanding of judicial systems and law enforcement, and they often smuggle people in and out of countries without getting caught [7]. This phenomenon poses a national security risk as many terrorist organizations fund themselves through human trafficking, and they can gain easy access to target countries through trafficking routes [7]. As long as there is a market for human trafficking, criminals, terrorists

A. Szakonyi (✉)

Public and Social Policy, Saint Louis University, Saint Louis, MO, USA

e-mail: annamaria.szakonyi@health.slu.edu

H. Chellasamy · A. Vassilakos · M. Dawson

Center for Cyber Security and Forensics Education, Illinois Institute of Technology, Chicago, IL, USA

and criminal organizations will continue making money from exploiting vulnerable populations.

In this age, data is the new currency as organizations use this in exchange for their services provided. However, as data is transformed into information, the possibilities become limitless. Previously, law enforcement organizations placed a significant work effort to uncover criminal activities such as human trafficking. Therefore, technologies can further aid in these law enforcement agencies in analyzing patterns. These patterns range from social media text, clustering, Dark Web photo metadata analysis for time and location.

It is necessary that data about human trafficking is analyzed using advanced data visualization and machine learning technologies. Through the investigation and analysis of data about human trafficking, researchers and law enforcement may gain better insight into its common patterns. In this research, the authors uncover patterns of human trafficking involving victim gender, type of victim exploitation, victim's country of citizenship, and country of exploitation. Additionally, by recognizing patterns in data from Blockchain activity, dark web marketplaces, or large human trafficking datasets, researchers can uncover hidden truths about human trafficking and become better prepared to fight the crime and rescue its victims.

64.2 Blockchain and Bitcoin

In recent years, cryptocurrency has been the rave, but that has been surrounding the exploding growth in Bitcoin's value. There are only aspirations and hopes of regulating cryptocurrencies [8]. In the meantime, this is the preferred form of payment for the exchange of illegal goods and services. One researcher takes an in-depth look into Bitcoin money laundering, exploring the negatives and positive outcomes of using cryptocurrency [9].

Even though the hype is slowing down around cryptocurrency, this is still the currency of choice to evade law enforcement [10]. Among the largest unregulated markets in the world are cryptocurrencies. Researchers estimate approximately \$76 billion of illegal activities per year, with one-quarter of Bitcoin users involved [11]. These numbers are astronomical and transforming the known black markets by enabling new e-commerce.

There has been a movement to trace criminal activity across the Bitcoin blockchain. By examining the blockchain activity through a process called clustering, discovering accounts purpose uncovers what type of storefront it is linked to. For example, if an account is used to make purchases on a Dark Web marketplace, we can begin to pinpoint appearances tied to the same Bitcoin wallet. This action may mean the same entity also controls them. Once that entity becomes known, then analysis can be done to begin uncovering who

that entity is through methods such as Open Source Intelligence (OSINT) and other forms of intelligence analysis coupled with data-driven tools.

64.3 Applying Machine Learning to Learn from Data

Unfortunately, due to the nature of this phenomenon, there are only limited datasets available publicly to study human trafficking. In addition, there seems to be an inconsistency in various types of data available from different organizations. There is no central authority that publishes an exhaustive set of data from various anti-trafficking organizations, combining data from not only rescue efforts but also from law enforcement and other sources. Therefore, it is extremely difficult to paint an all-encompassing picture of this phenomenon.

This study used the dataset from the Counter Trafficking Data Collaborative (CTDC), the "first global data hub on human trafficking, publishing harmonized data from counter-trafficking organizations around the world" [12]. This dataset "comprises identified or self-reported cases of victims of trafficking" [13].

The researchers conducted an exploratory data analysis, iteratively reviewing and interpreting the dataset. Some assumptions were made, such as the assumed relationship between gender and the type of exploitation (i.e. men are more likely to be exploited for certain types of labor trafficking), or some assumed relationships between countries of origin and type of exploitation (such as labor exploitation in the fishing industry in Thailand). Some high-level patterns were uncovered.

The researchers observed that the count of female victims was greater than the male victims (Fig. 64.1). The researchers observed that in most age groups, the female victims have a higher count. The single age group where male victims have a slightly higher count is the 39–47, with 1455 male victims over 1366 victims (Fig. 64.2).

In the data analysis, the Philippines was the country of citizenship with the highest number observed, with 11,365 victims reported to hold citizenship there (Fig. 64.3). Ukraine and the Republic of Moldova followed, with 7761 and 5901 victims, respectively. The USA and Cambodia concluded the top five countries, where 3636 and 1979 victims reported citizenship (Table 64.1). However, it is essential to mention the limitations in overall data, namely, that these datasets are provided voluntarily by organizations, and therefore, there is no equitable distribution of actual results from all countries and players. Even though some countries may have higher numbers in this dataset, that may not necessarily reflect an entirely accurate picture of the actual state of affairs, simply based on what data organizations provide to CTDC.

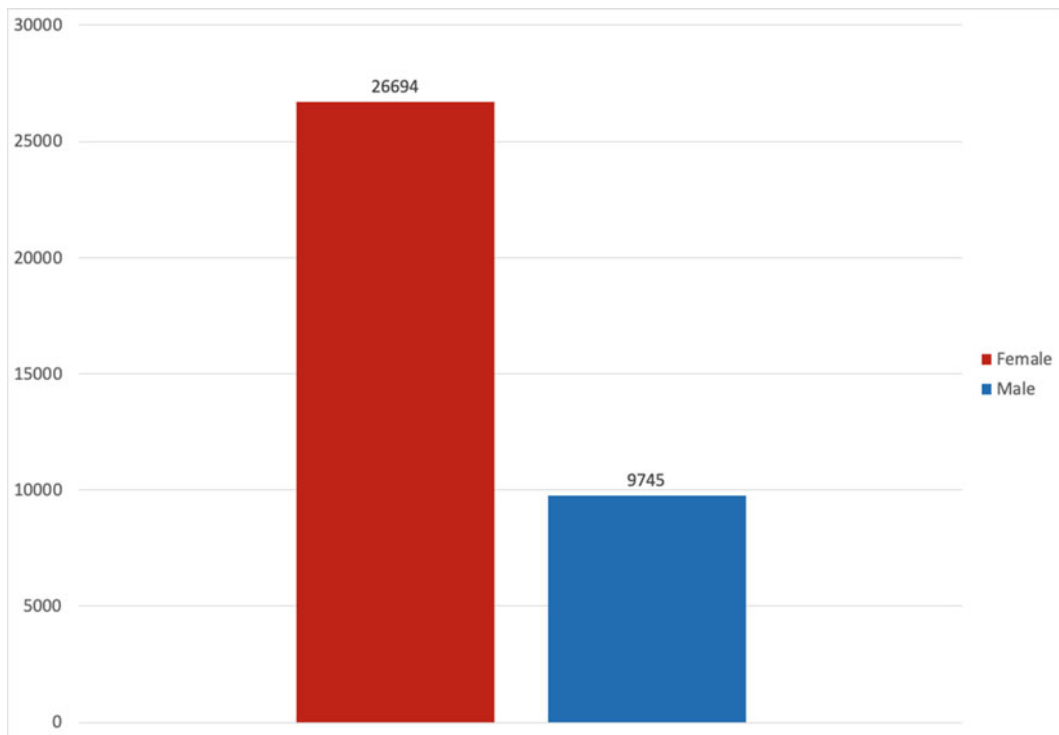


Fig. 64.1 Comparison between gender of victims shows an overall higher number of female victims

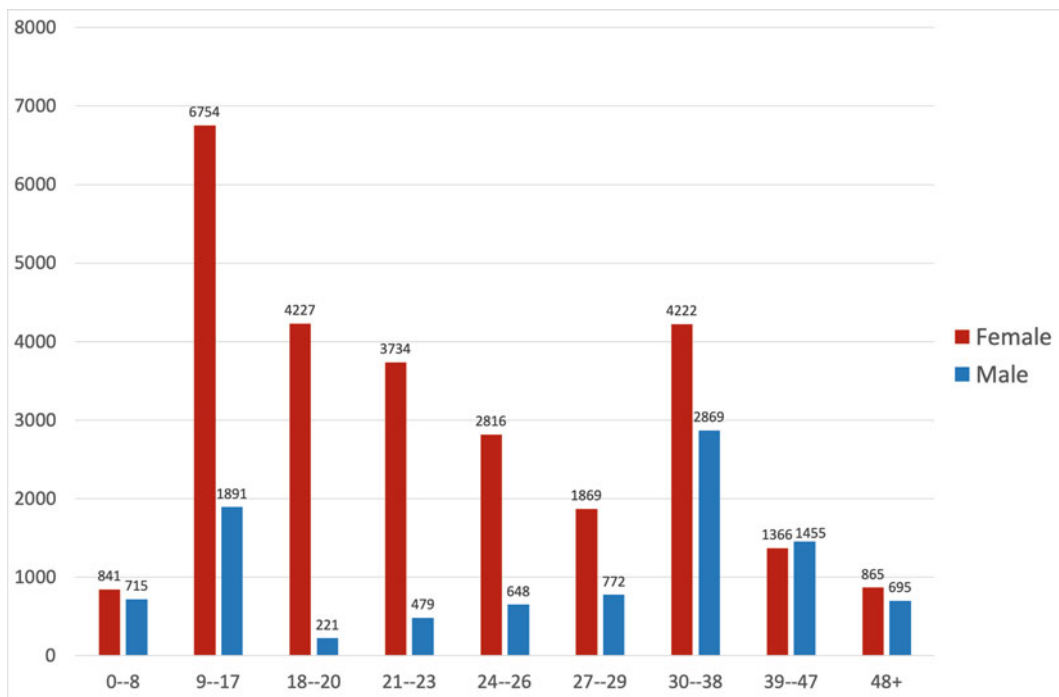


Fig. 64.2 Count of victims with age and gender filters shows one age group (39-47) with higher male victim counts

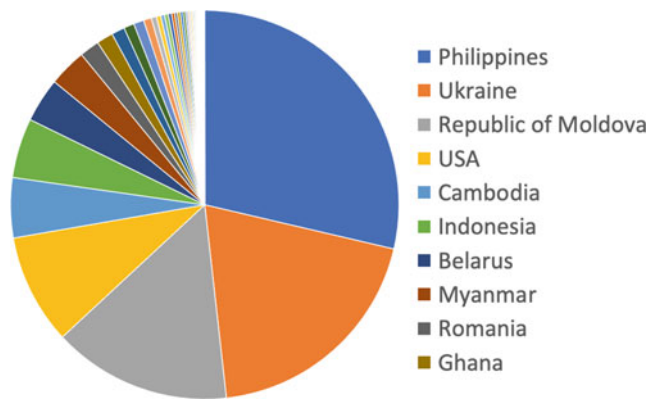


Fig. 64.3 Pie graph for victims' country of citizenship shows the Philippines as the country with most cases

Table 64.1 Top 10 countries of reported citizenship

Top 10 countries of citizenship	Amount
Philippines	11,365
Ukraine	7,761
Republic of Moldova	5,901
USA	3,636
Cambodia	1,979
Indonesia	1,971
Belarus	1,463
Myanmar	1,250
Romania	655
Ghana	544

For further analysis, the researchers reviewed patterns related to the types of exploitation and gender. Even though males get sexually exploited, the combination of sexual and labor exploitation as well as forced marriage did not show up as a male value. Therefore, it is clear that certain types of exploitation only exist for females.

When looking at country of origin (or assumed country of origin labeled as citizenship in the dataset) and country of exploitation, there is a clear pattern of movement present. The linear line can be observed for local cases, however, there is also a significant number of plots on the graph that show movement of victims from one country to another (Fig. 64.4). Mapping individual country movement could be an interesting topic to investigate with the help of geospatial data.

Next, the researchers conducted a regression analysis to uncover patterns and predict outcomes of human trafficking using the CTDC dataset. The research looked at gender differences, as well as exploitation differences of victims, namely, to understand differences between male and female targets and the types of exploitations victims may face. Therefore, both gender and exploitation types were chosen as

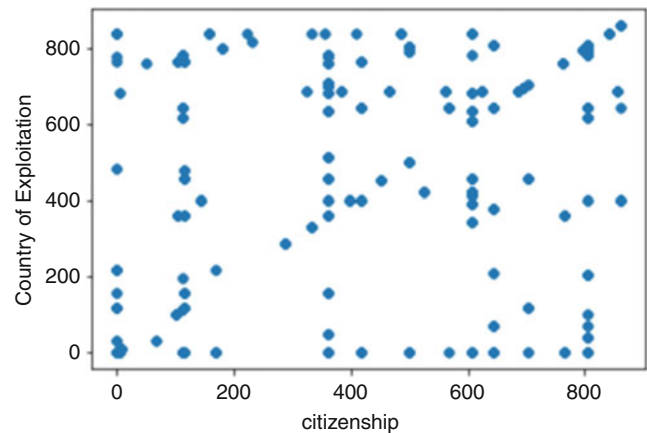


Fig. 64.4 Mapping country of origin to country of exploitation shows the local cases across the linear line on the graph, however, it is clear that there is a movement of victims from country of origin to the country of exploitation

output variables, and two separate models were implemented, one for each output.

When gender was defined as the dependent variable (the output), the focus was on how all the available features can explain the differences between what male versus female victims are targeted for. When exploitation was selected as the output, the question was focused on understanding the various tactics and the targeted victims of traffickers in order to uncover what categories of exploitation a victim would face, which can be useful in prevention and case prioritization efforts (i.e. if a victim is more likely to be sexually exploited when they are underage, that case should be prioritized).

After defining these outputs for each model, the dataset was split into train and test datasets (75–25% split) using the train-test split class of Scikit-learn. The authors implemented a decision tree to investigate classification of these outputs [14]. These results were further explored with random forest classifier and random forest feature importance to rank the various features and to understand what features may be better predictors.

Using decision tree model, gender can be predicted with an 83% accuracy on test data using all features after data cleaning. However, this makes the tree itself very complex and difficult to visualize. The more features were removed, the lower the accuracy became. Therefore, it seems like the complex combination of different variables is actually needed to provide accurate predictions for law enforcement and anti-trafficking organizations to aid them in the fight.

The accuracy with random forest classifier was 83% for gender. Random forest feature importance resulted in expected findings: features 24 (country of exploitation), 2 (citizenship) and 21 (type of exploitation) being the most important features in the prediction (Fig. 64.5).

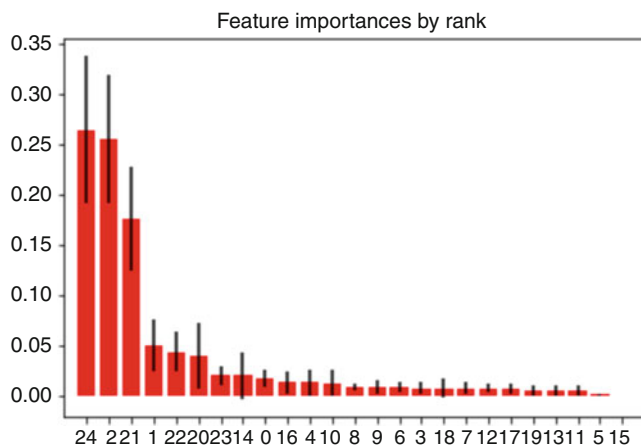


Fig. 64.5 Feature importance of random forest classifier for gender output shows features 24 (country of exploitation), 2 (citizenship) and 21 (type of exploitation) being the most important features in the prediction

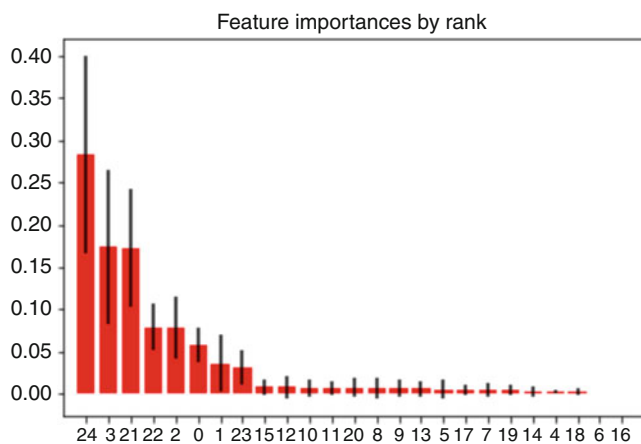


Fig. 64.6 Feature importance of random forest classifier for type of exploitation output shows features 24 (country of exploitation), 3 (citizenship), and 21 (means of control not specified) as the most important features for the prediction

For exploitation, with the decision tree algorithm, the results showed 88% model accuracy on test data. The accuracy on the random forest classifier was also 88%. The feature importance for exploitation identified features 24 (country of exploitation), 3 (citizenship), and 21 (means of control not specified) as the most important features for the prediction (Fig. 64.6). Means of control not specified is technically a labelled missing value describing that there was no data regarding how the victims were controlled, that column could have been dropped for even better predictions.

64.4 Future Research

Future research needs to be done on other Internet technologies, additional social media platforms, and devices that

collect data that could provide insights beyond what has been discussed in this paper. This would include looking at photo metadata and using it with applications that map the Geographical Information System (GIS) data. GIS can be used for further analysis related to trends in international trafficking efforts, namely, how victims are transported from one country to another. Text analytics is yet another area that would uncover messages written similarly by other human traffickers to understand the commonly used words to start tracking perpetrators' activities online.

64.5 Conclusion

In conclusion, information is the new cash as associations utilize this in return for their applications. Nonetheless, as data transforms into information, the conceivable outcomes become boundless. Beforehand, law enforcement associations set an immense work exertion to reveal crimes, for example, illegal exploitation. In this manner, advances can additionally help these law enforcement organizations in investigating designs. These examples range from online media text, grouping, Dark Web photograph metadata investigation for time and area. In this paper, the scientists provided a foundation regarding how illegal exploitation occurs, surveyed the utilization and ramifications of digitized money, and applied machine learning procedures to examine openly accessible datasets. The examination gave suggestions identified with information assortment, the executives, and investigation to help the indispensable battle against people's abuse.

An essential item learned is that international collaboration is crucial, so data provides a full global picture to enable global law enforcement efforts. This requires organizations to have staff aware of these Internet-based methods for exploitation and how one can use data science techniques to uncover patterns. Furthermore, while uncovering patterns of the exploitation, correlate the data to provide meaningful revelations that lead to an arrest. Understanding exactly how the illicit entities act on the Internet and beyond will allow for situational readiness in combating this phenomenon.

References

1. Department of Homeland Security (DHS) (ed.), What Is Human Trafficking? (2020, May 06). Retrieved October 09, 2020, from <https://www.dhs.gov/blue-campaign/what-human-trafficking>
2. International Labor Organization (ILO) (ed.), Forced labour, modern slavery and human trafficking. (2020). Retrieved October 09, 2020, from <https://www.ilo.org/global/topics/forced-labour/lang%2D%2Den/index.htm>
3. Oram, S., Abas, M., Bick, D., Boyle, A., French, R., Jakobowitz, S., . . . Zimmerman, C. (2016). Human trafficking and health: a survey of male and female survivors in England. *Am. J. Public Health*, 106(6), 1073–1078. doi:<https://doi.org/10.2105/ajph.2016.303095>

4. C. Zimmerman, M. Hossain, C. Watts, Human trafficking and health: a conceptual model to inform policy, intervention and research. *Soc. Sci. Med.* **73**(2), 327–335 (2011). <https://doi.org/10.1016/j.socscimed.2011.05.028>
5. J.H. Coverdale, M.R. Gordon, P.T. Nguyen, *Human Trafficking: A Treatment Guide for Mental Health Professionals* (American Psychiatric Association Publishing, Washington, DC, 2020)
6. S. Welch, Human trafficking and terrorism: utilizing national security resources to prevent human trafficking in the Islamic state. *Duke J. Gender Law Policy* **24**(165), 166–188 (2017)
7. R. Pati, Human trafficking: an issue of human and national security, *4U. Miami Nat'l Security & Armed Conflict L. Rev.*29 (2014). Available at: <http://repository.law.miami.edu/umnsac/vol4/iss2/5>
8. A. Narayanan, J. Bonneau, E. Felten, A. Miller, S. Goldfeder, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction* (Princeton University Press, Princeton, New Jersey 2016)
9. D. Bryans, Bitcoin and money laundering: mining for an effective solution. *Ind. LJ* **89**, 441 (2014)
10. R. Wolfson, Tracing illegal activity through the bitcoin blockchain to combat cryptocurrency-related crimes. (2018, December 15). Retrieved September 12, 2020, from <https://www.forbes.com/sites/rachelwolfson/2018/11/26/tracing-illegal-activity-through-the-bitcoin-blockchain-to-combat-cryptocurrency-related-crimes/>
11. S. Foley, J.R. Karlsen, T.J. Putniņš, Sex, drugs, and bitcoin: how much illegal activity is financed through cryptocurrencies? *Rev. Financ. Stud.* **32**(5), 1798–1853 (2019)
12. Counter Trafficking Data Collaborative, Telling Their Stories Through Open Data. (n.d.). Retrieved from <https://www.ctdatacollaborative.org/>
13. Counter Trafficking Data Collaborative, Data Codebook. (2017). Retrieved from https://www.ctdatacollaborative.org/sites/default/files/CTDC%20codebook%20v6_0.pdf
14. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**(85), 2825–2830 (2011) Retriever from <https://scikit-learn.org/stable/about.html#citing-scikit-learn>

Data-Driven Identification of Pedagogical and Curricular Factors Conducive to Student Satisfaction

Laura Sorto, Sourav Mukherjee, and Vasudevan Janarthanan

Abstract

In this paper, we apply machine learning methods to the problem of identifying key pedagogical and curricular factors that play a critical role in determining student satisfaction. Using standard end-of-semester course evaluations, we employ two different approaches to address this problem, namely, training a variety of regression models of overall satisfaction based on specific characteristics of teaching style and course contents, and computing correlations between overall satisfaction metrics and these specific characteristics. To validate our approach, we present a case study using data from a public institution of higher education. Based on empirical evidence, we identify a key subset of the questions included in course evaluations that are the biggest determinants of student satisfaction.

Keywords

Machine learning · Regression · Feature importance · Correlation · Higher education · Students' evaluations of teaching (SET) · Students' Evaluations of Educational Quality (SEEQ) · Faculty development · Curricular innovation · Education reform with technology

65.1 Introduction

Achieving a high degree of overall student satisfaction is critical to the success of both students and educational institutions. For example, the U.S. Bureau of Labor Statistics reports higher median incomes and lower unemployment

rates for those that have completed post-secondary degrees compared to those with only high-school diplomas [15]. Statistics Canada compares the incomes (cumulative over 20 years) of those with post-secondary degrees in various disciplines to the incomes of those with only high-school diplomas and reports that in several fields, the former category earns more [12]. Evidently, whether students remain engaged and motivated to complete their programs has pronounced economic ramifications. Moreover, [15] also shows that advanced degrees such as Master's and Doctorate degrees are associated with even higher median incomes and lower unemployment rates. Therefore, ensuring student satisfaction may significantly benefit students in the long term by inspiring them to pursue advanced education. From a university's viewpoint, student satisfaction may promote continued or increased enrollment, greater interest in advanced degrees, broader engagement in research, and more.

In this paper, we use machine learning based methods to identify key pedagogical and curricular factors that influence student satisfaction. Specifically, we employ two approaches, namely, (a) training predictive models of overall satisfaction based on specific characteristics of teaching style and course content, and (b) computing correlations between each of these specific characteristics and the metrics of overall satisfaction. In a case study using end-of-semester student satisfaction questionnaires from a public institute of higher education, we demonstrate that both approaches identify a small subset of the specific aspects that are key determinants of overall student satisfaction. We also show that choice of discipline does not have any appreciable influence on overall student satisfaction.

65.2 Related Work

The application of machine learning methods to various problems in higher education has emerged as an important area

L. Sorto · S. Mukherjee (✉) · V. Janarthanan
Fairleigh Dickinson University, Vancouver, BC, Canada
e-mail: laury98@student.fdu.edu; sourav@fdu.edu; v_janart@fdu.edu

of research in recent years. For example, student retention is crucial to the success of students pursuing their educational goals as well as that of institutions providing such education. This has motivated the development of predictive models to understand key factors influencing student attrition (i.e., dropping out), as well as to identify students who are at a risk of dropping out so that early intervention is possible [5, 6, 8, 9, 14]. Equally, student performance is also an important determinant of success, motivating research in machine learning models for predicting performance and identifying students at a risk of failing [2, 3]. Student satisfaction plays an important role in determining student retention and performance, and has been viewed as multidimensional by Marsh [10] and unidimensional by Abrami [1]. A widely adopted method for soliciting students' evaluations of teaching (SET) is the Students' Evaluations of Educational Quality (SEEQ) by Marsh [10]. Using machine learning methods, we show that a small number of questions asked in SEEQ are the biggest determinants of satisfaction.

65.3 Methodology

65.3.1 Preliminaries

In this paper, we assume a dataset $\mathbb{D} = \{(\mathbf{x}_i, \mathbf{y}_i) : 1 \leq i \leq n\}$ of i.i.d. instances where each instance $(\mathbf{x}_i, \mathbf{y}_i)$ arises from an end-of-semester survey of a course, aggregated (averaged) at the class level. An end-of-semester questionnaire typically includes questions on various *specific* aspects of the course and teaching methods the responses to which are solicited as a numeric rating. In this paper, we assume all ratings to be scaled to the range $[0, 1]$. If there are p specific questions, and q choices of discipline, then the input vector is given by $\mathbf{x} = [x_1, x_2, \dots, x_p, x_{p+1}, x_{p+2}, \dots, x_{p+q}]$ where $[x_1, \dots, x_p]$ contains responses to specific questions, and $[x_{p+1}, \dots, x_{p+q}]$ represents a one-hot encoding of the discipline. Similarly, the questionnaire also includes r dimensions of *overall* satisfaction, also expressed as ratings, which constitute the output vector $\mathbf{y} = [y_1, \dots, y_r] \in [0, 1]^r$.

65.3.2 Problem Definition

Our goals stated above may be formulated into the following two problem statements:

P1. Given a dataset $\mathbb{D} = \{(\mathbf{x}_i, \mathbf{y}_i) : \mathbf{x}_i \in [0, 1]^{p+q}, \mathbf{y}_i \in [0, 1]^r, 1 \leq i \leq n\}$, find a function $f : [0, 1]^{p+q} \rightarrow [0, 1]^r$ that maps input vectors to output vectors ($f(\mathbf{x}) = \mathbf{y}$) such that f minimizes the expected error with respect to a predefined loss function over the underlying data distribution from which \mathbb{D} was sampled.

P2. Identify which features of the input vector $\mathbf{x} \in [0, 1]^{p+q}$ have the highest influence in determining the value of the output vector $\mathbf{y} \in [0, 1]^r$.

65.3.3 Approach

Figure 65.1 depicts the architecture of our framework for training predictive models of overall student satisfaction as well as for identifying the most relevant contributing factors. Our framework accepts as input a collection of PDF documents containing students' end of semester evaluations, aggregated at the class level. It then performs the following operations on the data.

Data Extraction The text of the PDF documents is programmatically converted to a table where every row corresponds to a single instance of a course offering. The columns of the table include answers (ratings) to the questions of the corresponding end-of-semester survey aggregated at the class level, as well as the discipline associated with the course.

Data Preprocessing Ratings are scaled to $[0, 1]$, whereas choice of discipline is represented using one-hot encodings.

Multi-Output Regression Since the outputs of the predictive models are numeric ratings, we adopt the approach of training multi-output regression models using the following algorithms: 1. K -nearest neighbors regression, 2. linear support vector regression, 3. support vector regression with radial basis function (RBF) kernel, 4. linear regression, 5. decision tree, and 6. random forest.

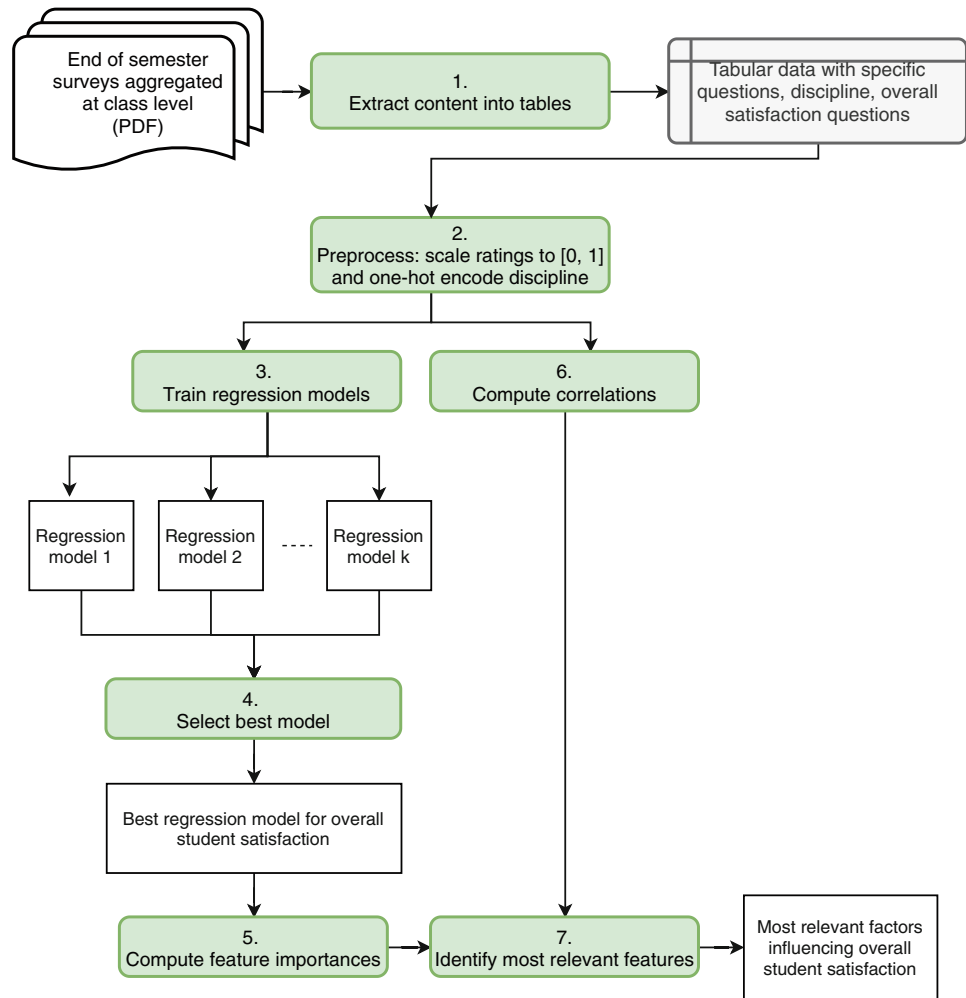
Model Evaluation and Selection To evaluate our models, we use explained variance and mean squared error, measured using k -fold cross validation. Given a multi-output regression model $[y_1, \dots, y_r] = f([x_1, \dots, x_p, x_{p+1}, \dots, x_{p+q}])$, we compute the explained variance of the model over a validation dataset as follows:

$$expl_var(f) = \frac{1}{r} \sum_{i=1}^r \left(1 - \frac{Var(y_i - y_i^{\text{pred}})}{Var(y_i)} \right)$$

where $\mathbf{y} = [y_1, \dots, y_r]$ is the true value of the output and $\mathbf{y}^{\text{pred}} = [y_1^{\text{pred}}, \dots, y_r^{\text{pred}}]$ is the output predicted by the model, and $Var(\cdot)$ denotes variance computed over the validation dataset. We compute the mean squared error of the model over the validation set using:

$$mse(f) = \frac{1}{r} \sum_{i=1}^r E((y_i - y_i^{\text{pred}})^2)$$

Fig. 65.1 Architecture for training predictive model of overall student satisfaction based on specific questions and discipline, and for identification of most relevant contributing factors



where $E(\cdot)$ denotes mean computed over the validation set. These metrics allow us to select the model best suited for relevant feature identification as described next.

Feature Importance Calculation Having selected the predictive model with the highest cross-validation score, we compute the importance of each feature in determining the output of this model. In our case study described in Sect. 65.4, random forest regression emerged as the model with the highest cross-validation score. Therefore, we use Gini importance [4] to rank features by relevance.

Correlation Calculation We also compute Pearson's correlation coefficients between input and target features and use the magnitudes of these correlations to determine feature relevance.

Identification of Most Relevant Features Finally, we take the intersection of relevant features obtained using Gini im-

portance and those obtained using correlations, resulting in a set of features identified as most relevant.

65.4 Evaluation

Dataset The dataset for our case study is derived from publicly available SEEQ survey results at the University of Maryland Baltimore County, USA, for courses offered from Fall 2016 to Fall 2019 semesters, available at <https://oir.umbc.edu/university-data/sceq-profiles/>. After discarding surveys with missing data, and selecting courses from the top 10 disciplines by enrollment, the resulting dataset contains results for 6961 offerings. We define the *target* $\mathbf{y} = [y_1, y_2, y_3]$ to be the answers (scaled to the range $[0, 1]$) to the questions categorized under *overall*, which are numbered 30, 31, 32 in the questionnaire. We define the *input* to our models to be of the form $\mathbf{x} = [x_1, \dots, x_{29}, x_{30}, \dots, x_{39}]$ where x_1, \dots, x_{29} are answers to the questions under every category except *overall* in the questionnaire, normalized to the range

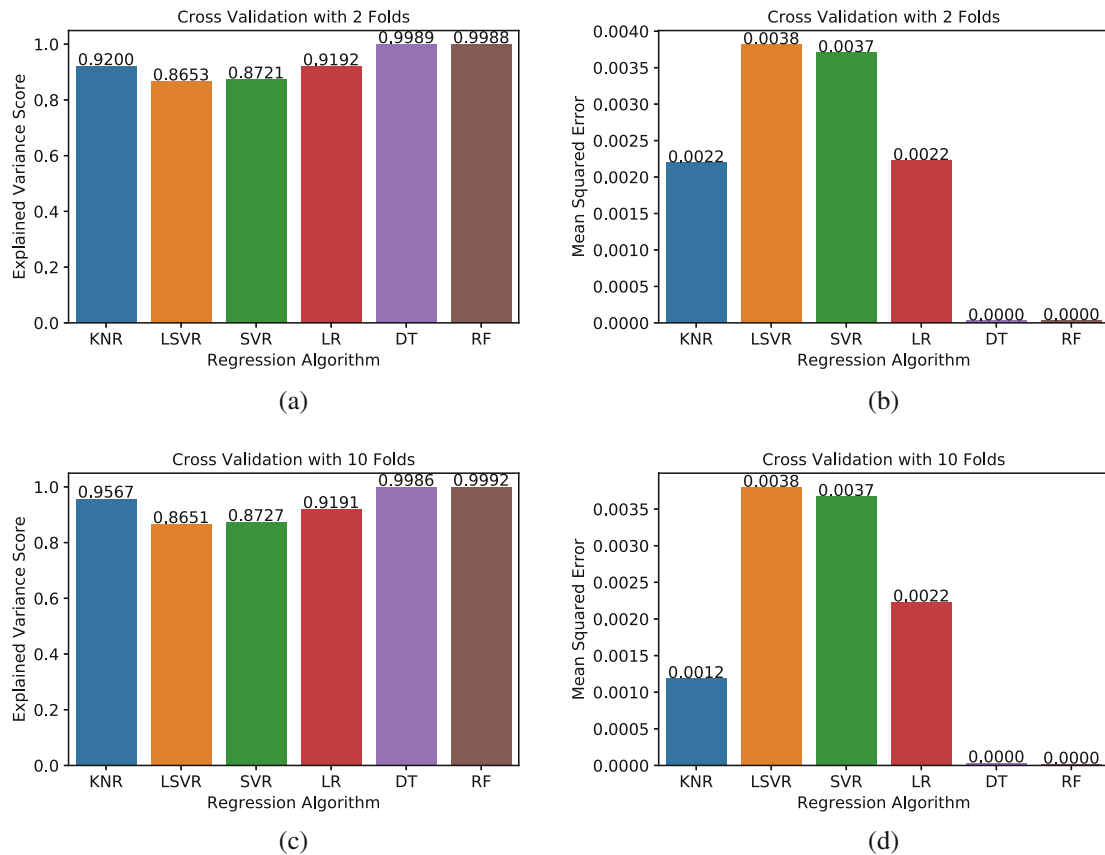


Fig. 65.2 Explained variance and mean square error metrics of the multi-output regressors using 2-fold and 10-fold cross validation. (a) Explained variance with 2-fold cross validation. (b) Mean squared error

with 2-fold cross validation. (c) Explained variance with 10-fold cross validation. (d) Mean squared error with 10-fold cross validation

[0, 1], whereas x_{30}, \dots, x_{39} represent one-hot encodings of the top 10 disciplines ranked by the number of students enrolled.

Experimental Setup After extracting data from PDF documents using PyPDF2,¹ we implement multi-output regression models using standard Python libraries [7, 11, 13, 16].

Results

Predictive Accuracy. Figure 65.2a and b show the 2-fold cross validation scores of the explained variance metric and mean-squared error for each of the above regression models. Similarly, Fig. 65.2c and d show the corresponding metrics with 10-fold cross validation. In view of the above results, we identify random forest regression as the best suited model for this predictive task, among the models tested in this study.

Feature Relevance. Figure 65.3 shows the Gini importance scores of the input features with respect to the targets (only the top 10 input features are shown). Figure 65.3a, b, and c

show the importance scores with respect to Questions 30, 31, and 32 (i.e., those belonging to the *overall* category in the questionnaire) taken individually, whereas Fig. 65.3d shows the Gini importance averaged over these three questions.

Correlations. The correlations of Questions 1–29 to each of Questions 30, 31, and 32 are shown in Fig. 65.4. Choice of discipline is not included in the plot since the highest observed correlation of any discipline choice with any of the target questions is 0.1337.

65.5 Discussion

In our evaluation, Figs. 65.3 and 65.4 both suggest that a small number of specific pedagogical and curricular traits contribute most heavily to overall satisfaction whereas other aspects probed by the questionnaire have much lower impact on satisfaction. Based on Fig. 65.3d we identify the top 5 features most relevant to overall satisfaction (see Table 65.1). Other questions in the survey have much lower impact, and choice of discipline has no appreciable impact, on overall satisfaction. A new faculty member, or one seeking to improve his/her teaching effectiveness, may therefore benefit most

¹<https://pypi.org/project/PyPDF2/>.

Fig. 65.3 Features ranked using Gini importance scores with respect to Questions 30, 31, and 32 taken individually as well as taken together (averaged). The importances are computed using a random forest regressor. Only top 10 features are shown in each case. (a) Gini importance scores with respect to Question 30. (b) Gini importance scores with respect to Question 31. (c) Gini importance scores with respect to Question 32. (d) Gini importance scores averaged over Questions 30, 31, and 32

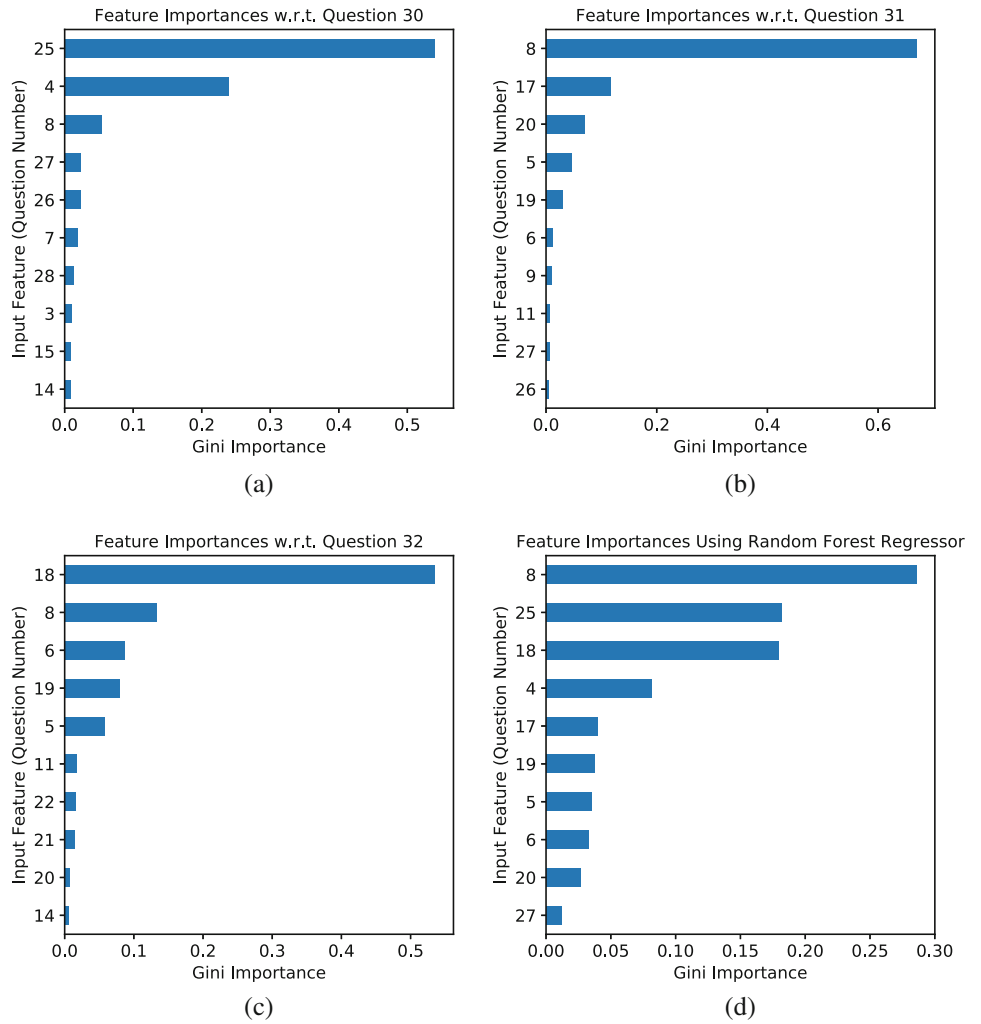


Fig. 65.4 Correlations of Questions 1–29 to Questions 30, 31, 32

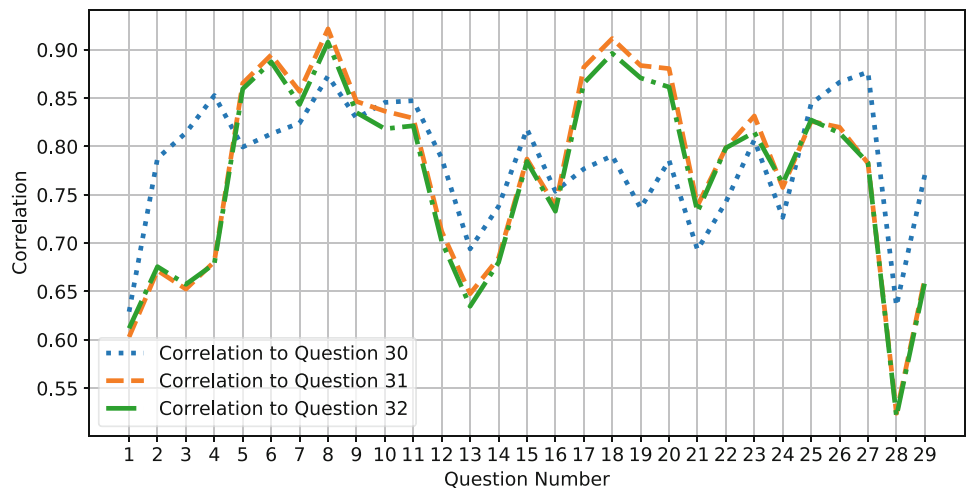


Table 65.1 Top 5 features most relevant to overall satisfaction based on Gini importance

#	Question
8	The instructor's style of presentation held my interest during class.
25	Feedback on examinations/graded materials was valuable.
18	The instructor made students feel welcome in seeking help/advice in or outside of class.
4	I have learned and understood the subject materials of this course.
17	The instructor was friendly towards individual students.

from focusing on an engaging presentation style, providing useful feedback on evaluations, being approachable to students, ensuring that students comprehend the material, and maintaining a friendly disposition.

65.6 Conclusion

To summarize, we have used machine learning methods to identify key determinants of student satisfaction. Based on empirical evidence from a case study, we have identified a small subset of the pedagogical and curricular aspects covered in the Students' Evaluations of Educational Quality (SEEQ) questionnaire that are most influential in determining overall student satisfaction. Other aspects probed in the questionnaire, and choice of discipline, have not demonstrated significant influence overall student satisfaction. In view of the above findings, we have identified key pedagogical attributes that faculty members may prioritize in order to improve teaching effectiveness.

Acknowledgments We thank Dr. Tim Oates (University of Maryland Baltimore County, USA) for pointing us to the dataset used in this paper.

References

1. P.C. Abrami, How should we use student ratings to evaluate teaching? *Res. Higher Educ.* **30**(2), 221–227 (1989)

2. M.M. Alam, K. Mohiuddin, A.K. Das, M.K. Islam, M.S. Kaonain, M.H. Ali, A reduced feature based neural network approach to classify the category of students, in *2nd International Conference on Innovation in Artificial Intelligence* (2018), pp. 28–32
3. V.K. Anand, S. Rahiman, E.B. George, A.S. Huda, Recursive clustering technique for students' performance evaluation in programming courses, in *2018 Majan International Conference (MIC)* (2018), pp. 1–5
4. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees* (Wadsworth, Monterey, 1984)
5. K.E.K. Chai, D. Gibson, Predicting the risk of attrition for undergraduate students with time based modelling, in *CELDA* (2015)
6. D. Delen, A comparative analysis of machine learning techniques for student retention management. *Decis. Support Syst.* **49**(4), 498–506 (2010)
7. J.D. Hunter, Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* **9**(3), 90–95 (2007)
8. J.-W. Jia, M. Mareboyana, Predictive models for undergraduate student retention using machine learning algorithms, *Transactions on Engineering Technologies* (Springer, New York, 2014), pp. 315–329
9. I. Lykourantzou, I. Giannoukos, V. Nikolopoulos, G. Mparidis, V. Loumos, Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ.* **53**(3), 950–965 (2009)
10. H. Marsh, Seeq: a reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *Br. J. Educ. Psychol.* **52**, 77–95 (1982)
11. W. McKinney, Data structures for statistical computing in python, in *9th Python in Science Conference* (2010), pp. 51 – 56
12. Y. Ostrovsky, M. Frenette, The cumulative earnings of postsecondary graduates over 20 years: results by field of study. <https://www150.statcan.gc.ca/n1/pub/11-626-x/11-626-x2014040-eng.htm>. Accessed 7 June 2020
13. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
14. S. Ram, Y. Wang, F. Currim, S. Currim, Using big data for predicting freshmen retention, in *International Conference on Information Systems - Exploring the Information Frontier, ICIS* (2015)
15. U.S. Bureau of Labor Statistics. Unemployment rates and earnings by educational attainment. <https://www.bls.gov/emp/chart-unemployment-earnings-education.htm>. Accessed 7 June 2020
16. S. van der Walt, S.C. Colbert, G. Varoquaux, The numpy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **13**(2), 22–30 (2011)

Joseph Elliot and Daniel Berleant

Abstract

Information Quality (IQ) of a university website plays a major role in the decision process for prospective students when selecting a university for their higher education. Furthermore, current students and others rely on university websites for many other purposes. In this paper we identify university website information quality dimensions relevant to prospective and current students and other site users. We discuss the rationales for identifying these IQ dimensions and propose a University Website Information Quality (UWebIQ) framework to quantify the individual IQ dimensions as well as a strategy for defining the composite IQ for such a website. The outcome of this research is expected to provide insight for universities that wish to maximize the fitness for use of their websites.

Keywords

Composite IQ · Data quality · Dimension · Dimensional IQ · Enrollment · Information quality · Internet · WebIQ · World wide web · Website

66.1 Introduction

Information Quality (IQ) describes an information product's fitness for use for a specific task [1]. Yet this fitness is itself often a composite of a number of different, more specific dimensions. Each contributes to the overall infor-

mation fitness. How to calculate an overall quality (fitness) from ratings along these disparate dimensions is important because a single, summarizing quality rating is often needed. In addition to multiple ratings along different dimensions there may be multiple ratings of the same dimension. This problem is somewhat similar to the need to calculate a single overall quality rating using multiple estimates. Both of these problems present a combination of evidence challenge, a challenge that can be addressed using the approach which we present here.

Information quality (IQ) theories and frameworks have been increasingly researched, practiced and standardized at major government and non-government entities in order to assess, monitor and improve the quality of their information products [2, 3]. United States government agencies have enacted information quality guidelines [4]. For example, the Department of Health and Human Services (HHS) has published web standards required for the design and development of all HHS and priority websites [5]. The Energy Information Administration (EIA) produces and publishes standardized energy-related data such as world crude oil reserves by regions and countries, and relies on well-developed information quality techniques. In the education industry, however, universally accepted standards for university websites and for their in-depth analysis, assessment, and quantification with respect to their information quality is currently lacking. Thus, there is a need for web information quality frameworks for quantitatively assessing the qualities of university websites.

In the education industry, university and college websites should support efforts to recruit and retain students, reduce costs for staff to provide information to stakeholders individually and manually, lower costs related to physical printing, support pedagogical, research, and administrative functions, and so on. Fitness for use of university websites implies support for all of those functions and more. As one example, unavailable or inconsistent website links and low usability user interfaces could lower new student applications

J. Elliot (✉)
Rutgers University, Newark, NJ, USA
e-mail: joseph.elliott@rutgers.edu

D. Berleant
University of Arkansas at Little Rock, Little Rock, AR, USA
e-mail: jdberleant@ualr.edu

and enrollments. To characterize fitness for use of college and university websites, we describe a new information quality framework for this purpose based on accepted theories and foundations in the information quality domain. Using this proposed framework, it is possible for colleges and universities to design and implement upgrades of their websites based on objective computation of the information qualities along key dimensions and overall based on information qualities along all the dimensions.

66.2 Literature Review

Measuring information quality often requires combining different dimensions of quality which are all relevant in different ways to an overall quality assessment. Stvilia et al. [6] proposed a “general framework for IQ assessment.” They explained that aggregation, or clumping together of different things, refers in the information quality field to both grouping different but related information entities together, and combining measurements of information quality to get an overall quality rating. The latter type of aggregation is more relevant to the present report. They advocate a process that is complicated, involving concept trees, metrics, measurement representations, and IQ activity system value structure. A key subgoal of one approach is to identify and combine IQ value curves using transfer functions. An alternative approach is based on statistical profiles of IQ user evaluations. Two case examples both use factor analysis. Ultimately the paper provides a general framework but without specifying a particular algorithm for combining IQ measurements, which is a gap we and others have sought to bridge.

Robbins and Austin [7] advocate a process that contrasts with that of Stvilia et al. by avoiding the complexities of a multi-faceted process which has no clear step by step procedure. Their method combines information quality ratings on different dimensions by simply multiplying them. This was a reasonable model for the specific dimensions they chose, which were importance and completeness of the information item. This illustrates a model based on the specific meanings of stated dimensions. However, multiplying the information quality measures does not generalize well to arbitrary dimensions. It is ad hoc, and thus not suitable as a general model of combining information quality dimensions.

Robbins and Austin also combine quality ratings by an averaging process. This may be not unreasonable for combining multiple assessments of the quality of one information item. However it is less defensible for combining ratings on multiple dimensions for one information item because a quite low rating on an important dimension carries considerable weight that cannot be easily compensated for by high ratings on other dimensions. For example, if timeliness is an important dimension and the information is not timely, its

quality is significantly affected even if it rates highly on other dimensions. Thus there is a need for combination methods that circumvent this issue.

Quarati et al. [8] use an Analytic Hierarchy Process based approach to combine quality ratings of different dimensions in stages. First, dimension measurements are combined to give a quality rating to a category of dimensions, and then the category ratings are combined to give an overall rating. The method generalizes to different hierarchies, such as combining subdimension ratings to get a dimension rating, then dimensions to get a dimension category rating, then categories to get an overall rating. Each rating to be combined with other ratings is given a weight determined by crowd sourcing expert judgments, and the weighted ratings are averaged. Although, as we will see, simple averaging is problematic, the approach is certainly adaptable to other combination methods as well. It is suitable in general to problems in which a hierarchical organization of the combination process is indicated.

66.2.1 Identifying Information Quality Dimensions

Tao et al. [9] describe a process of identifying information quality dimensions based on user studies. They focus on the problem of identifying information quality dimension indicators, or attributes, which can be used to assess dimensions. They call these indicators drivers, and define a driver as an attribute of an information quality dimension that is perceived by users as indicating quality. In their investigation into the domain of health information websites, there is often ambiguity about whether a driver is really a dimension or vice versa. Additionally, they found that some drivers indicate the level of quality on more than one dimension. However, they did not address combining dimensions to produce a composite, overall information quality rating.

While information quality dimensions can be identified on a problem-dependent basis, another approach is to use general purpose dimensions that apply to a wide variety of problems and are defined by standardized references. An example of this approach is the twenty-first Century Integrated Digital Experience Act of 2018 (e.g. [digital.gov/resources/21st-century-integrated-digital-experience-act](https://www.digit.gov/resources/21st-century-integrated-digital-experience-act)). This standard applies to the design and development of all HHS/OS and priority websites [5]. The Act requires Federal Agencies websites to satisfy certain quality dimensions [10]. Specifically, they must: (1) be accessible to people with disabilities, (2) be consistent in appearance, (3) be authoritative in that they avoid redundancy with other websites, (4) have a site search capability, (5) have appropriate security, (6) “be designed around user needs with data-driven analysis,” (7) be customizable to user preferences, and (8) work properly

on mobile devices. A related standard provides a checklist of requirements for federal websites and digital services via the digital.gov site, which endorses guidelines and standards including from ref. [11].

Another influential set of general purpose information quality dimensions is Morville's User Experience Honeycomb, which defines a set of dimensions in order for information to provide a meaningful and valuable user experience [12]. These dimensions are posed as requirements for websites, specifically: (1) useful, (2) usable, (3) desirable, (4) findable, (5) accessible, (6) credible, and (7) valuable. These are adopted for example by Semantic Studios [13]. In contrast to the aforementioned twenty-first Century Integrated Digital Experience Act of 2018 standard, we have found that the Honeycomb can be applied to a college or university website setting without modification because redundancy with other websites (dimension 3 of [5]) is less of a concern.

Some dimensional frameworks have been studied specifically with respect to college and university websites. For example Mentes and Turan [14] focus on usability evaluation of the website of a particular institution, Namik Kemal University. The authors survey previous papers on assessing the website usability of specific universities. Given their focus on usability they need a framework that provides dimensions of usability. They list QIS, SUMI, NIST Web Metrics, MUMMS, and WAMMI, finally choosing WAMMI (**Website Analysis and Measurement Inventory**) for their study. Clearly usability is an important aspect of the fitness for use and thus the information quality of college and university websites. However usability alone is too limited to be the sole focus for assessing college and university website quality. That is why we use a more comprehensive framework based on Morville's User Experience Honeycomb.

66.2.2 Measuring an Information Quality Dimension

A related problem is combining multiple measurements of the same quantity to get an improved estimate of the quantity. Well established formulas exist for estimating the error and value of a metric composed of multiple measurements each with its own error [15]. The method gives the error in an estimate of a quantity as the square root of the sum of the squares of the errors of the components which compose it. Put another way, it applies the formula of the Pythagorean theorem. This approach could potentially be applied in estimating the information quality of a university website. For example, assume the true information quality for a dimension is Q , and two estimates or measurements of the dimension are Q_1 and Q_2 with errors e_1 and e_2 . Then the method estimates the value

of Q as $(Q_1 + Q_2)/2$ and the error as $+\sqrt{(e_1^2 + e_2^2)/(N-1)}$ where N in this case is 2.

The problem with applying such a method to information quality ratings is that the error or uncertainty in a rating of information quality along a given dimension is difficult to determine. Yiliyasi [16] proposes a somewhat complex but probabilistically well motivated approach of addressing this problem of combining information quality. That method was described for the problem of combining expert estimates of the same value, where each expert estimate has its own information quality. The method may be adaptable to the problem of combining assessments of the information qualities of different indicators of a given dimension. These indicators have also been called aspects, attributes, indicators, drivers, quality markers and surrogates (e.g. Tao et al. [9]). However combining indicators of a given dimension is not the same problem as combining different dimensions to give a high level or overall information quality rating, as required by the information quality framework described in the next section.

66.3 Design of an Information Quality Framework

Information can be defined as the product of an information system that processes raw data into a usable product that adds value for the information consumer [1]. Accordingly, we consider a university website an information product whose quality we wish to quantitatively assess as a composite of measurements along multiple dimensions.

We have selected IQ dimensions that are important to the quality of university websites from the perspective of their users, especially prospective and current students, and seek to quantify them in a transparent and reproducible manner. The rationale for using each of these IQ dimensions is discussed in detail in the following sections along with their application to university websites. However, it should be noted that there are other possible IQ dimensions such as *Trendiness*, *Attractiveness*, *Interaction*, *Theme*, etc., that might also be useful to include in a framework of this type if the practical difficulties in assessing and quantifying them can be solved. We identify the following IQ dimensions for our UWebIQ Framework. This list of seven factors, given at usability.gov, is the same as that proposed by Semantic Studio. That is because both are based on Morville's User Experience Honeycomb, which has those same seven factors. Building on that foundation of Morville's Honeycomb, we develop it further with quantitative metrics that help assess and improve the quality of university websites.

Here are the seven dimensions and the measurement process for each.

66.3.1 Useful

By useful is meant that content should be original and fulfill user needs. This may be quantified as follows.

- (a) Universities may determine how much money they spend on answering student and prospective student queries by email or phone (excluding communications with their instructors). The cost might, for example, be determined from salaries, time sheets, and/or budgets. A website that reduced this cost would have a higher quality. Since the cost is higher for *less* useful websites, it needs to be massaged so it is instead higher for *more* useful websites. An easy way to do that is to subtract it from 1.

The figure should also be normalized to a defined scale, here 0–100, so that it can be more easily compared and combined with other measures of information quality. A formula that meets these requirements is:

$$100 * \left(1 - \frac{\text{Cost of answering queries by humans}}{\text{Total technology infrastructure cost}} \right).$$

This is based on ongoing cost of answering queries as a fraction of the total ongoing technology cost. Like any fraction of a total it has values between 0 and 1. That fraction is inverted by subtracting it from 1, then expanded to the range [0, 100]. To evaluate website improvements with respect to this metric, the formula can be recalculated periodically. If its value increases, that means the institution had benefited financially, validating the changes as improvements in website quality.

- (b) As an alternative, selected web pages could require a viewer to answer a single question about usefulness in order for the viewer to be permitted to follow a link from the page or submit a form. The response to the question is scored on a 0-to-100 range and the average over all users answering the question is the usefulness rating of the page. The overall site usefulness rating is the average of the page ratings over all assessed pages.

Higher values indicate better user ratings, validating the metric.

- (c) As a third option, uses of online real time help chat windows could be counted, as well as the fraction of those uses that result in providing a website link. The site is more useful if a link can be provided to resolve a chat session. This option may be applied to a website if it has help chat session functionality. The resulting count may be normalized by multiplying by 100. The formula is:

$$100 * \left(\frac{\text{Number of help chat sessions resulting in a link to a website page}}{\text{Total number of help chat sessions}} \right)$$

The range of the fraction in this formula is from 0 (indicating help sessions did not result in links, hence a low information quality for the site) to 100 (all help sessions resulted in a link). Higher values indicate more a more effective business process, likely due to a better website but also potentially due to better employee training or other non-website activities.

Some universities may choose to implement 2 or 3 of the above three options. This would require combining the 2–3 metrics into a single measure of usefulness. The mean of the metrics would work for this [18].

66.3.2 Usable

By usable is meant that the website is easy to use. To quantify this, universities can determine how many queries by email and telephone they get that are answered on web pages, relative to the total SSCHs (student-semester credit hours) or other measure of university instructional load. This can be done by requesting the relevant employees to tabulate the number of such inquiries they deal with. They can be incentivized to do this by pointing out that the information will be used to make the website more usable in order to reduce the number of emailed and called-in inquiries they have to deal with. A formula that can model this is:

$$100 * \frac{\text{SSCHs}}{\text{queries} + \text{SSCHs}}.$$

This formula has the value 100 if there are no such queries, suggesting all web pages are easily found and the website is sufficiently usable. On the other hand, the formula descends toward the value 0 when there is a very large number of queries, suggesting the website has poor usability.

Alternatively, real time chat window uses that result in providing a link to a web page could be counted, since the site would be more usable if the user could find the pages themselves. A formula based on this approach that would work is:

$$100 * \frac{\text{SSCHs}}{\text{chat window uses resulting in a link} + \text{SSCHs}}.$$

This has values approaching 100 when there are very few such chat window uses, suggesting a highly usable website where people can readily find the pages they need. The

value declines for more chat window inquiries resulting in a page link, suggesting a website with pages containing the information people need, but they are not finding those pages.

If both metrics are available, then they may be averaged. Improvements in this metric represent more effective communication to users from the website, thus validating the metric.

66.3.3 Desirable

The content of the website should evoke users' positive emotions and sense of appreciation so that users desire to use the website [12]. A standard approach to this that students are already familiar with is the 1–5 star rating system. It is commonplace on the web and could be applied here. In one approach, web pages on a university website could each end with a brief request for the reader to click 1 to 5 stars. This system is readily implemented because third party providers make it easy to add this to a web page. An example service is at rating-widget.com. The overall rating of the site would be based on a weighted average of the ratings of the individual pages. The weight would be based on the number of ratings a page has. More ratings should increase the weight of the overall rating of a page, since a heavily rated page is probably a heavily used page, and the desirability of a page is more important the more frequently it is visited. The simplest way to meet this requirement is just to add up all the ratings regardless of what page any rating is for, then divide by the total number of ratings. The rating for the website would then be:

$$S = \frac{\sum_{i=1}^p \sum_{j=1}^{n_i} r_{i,j}}{\sum_{i=1}^p \sum_{j=1}^{n_i} 1}$$

which says, for each of the p pages, add up its n ratings r , then add up all of the p subtotals. Finally divide by the total number of ratings across all the pages. Note that each $r_{i,j}$ will be a number between 1 and 5 (because of the 1-to-5 star rating system). Normalizing to the interval from 0 to 100 is thus done by mapping the interval of possible values of S , which is [1, 5], to [0, 100]:

$$\text{Desirability} = 100 * \frac{S - 1}{4} .$$

66.3.4 Findable

A website is findable to the degree that web search engines list it with a ranking that brings it to the attention of people querying. To measure this, use a major search engine, such as google.com. Determine a set of queries that are believed to

be representative of those used by people who the university wishes to reach through web search queries. The findability will be the average ranking of the university's website over those queries. This will be useful for testing whether new SEO (Search Engine Optimization) actions embedded in the website are working better than previous ones. To implement this, a university can:

1. determine a set of queries and whether it would be reasonable, for each one, to hope it would be the top hit on a specified search engine (e.g. google.com), or that it at least be in the top 10 hits;
2. check each query to see if it meets the stated requirement (top hit vs. top 10 hits);
3. score each query at 1 if it meets its requirement or 0 if it does not;
4. let $q = 100 * (\text{average of the scores of the queries})$.

Then q is the information quality of the website along the *findable* dimension. Improvements in this metric represent a clearly desirable change to the findability of the website, validating the metric.

66.3.5 Accessible

There are various standards that exist for this and can serve as the foundation of a metric. Semantics Studio's definition of the *Accessible IQ* dimension is that a website should be accessible to people with disabilities [13]. Usability.gov refers to *Accessible IQ* as *Accessibility IQ*, which focuses on how people with disability would access or benefit from a website [12]. Lee et al. [17] define *Accessibility IQ* as the ease at which an information product is retrievable, obtainable and quickly accessible when needed [21]. The Bureau of Internet Accessibility (BOIA) provides a free website accessibility compliance scan [18]. BOIA [19] endorses the W3C's WCAG [11].

Many websites are not compliant with accessibility standards. A common example is that mobile websites do not always provide the same interface and links as the standard desktop interface.

66.3.6 Credible

The Stanford University Web Credibility Research site lists 10 guidelines for website credibility [20]. These are described next with how they may be measured for the present web quality framework.

- (1) *Information accuracy should be easily verified.* We wish to measure this in a way that is readily measured yet

is also a good proxy for information accuracy. To this end, recall that scholarly publications use citations interspersed within text as the classical method for providing verifiability of statements. The analog of this in a hypertext context, of which the web is the main example, is links embedded in the content of a web page. It is relatively straightforward to count the fraction of pages on a website that contain such links. A formula that naturally ensues as a metric scaled from 0 to 100 is:

$$100 * (\text{Fraction of pages on a site with link (s) embedded in their content}).$$

- (2) *Give the organization of the website.* One standard way to show website organization is to show a horizontal navigation menu at the top of the page. Other standard methods include hamburger menus, sidebars, fat footers, and breadcrumbs (hierarchical indicators showing the depth and location of a page within the website). All of these are popular web page design constructs (e.g. [21].) and are readily transformable into metrics by simply determining the fraction of web pages on a site that include one or more of these website organization techniques. Scaling from 0 to 100 gives the formula:

$$100 * (\text{Fraction of pages on the website showing website organization}).$$

For some sites this will be near or at 100, since many sites will have all pages containing a standard site organization element. On the other hand, some sites might score very low.

- (3) *Display the content expertise behind the website.* One approach to determining the degree to which expertise is highlighted is to check whether there are page footers giving information on the expertise of contributor(s) to the page. Statements of expertise are more credible if the contributor(s) with the expertise are identified by name. This leads to the following formula, which is scaled from 0 to 100:

$$100 * (\text{Fraction of pages with footers giving the expertise of at least onedirectcontributor to the page content}).$$

A simple statement like “This content was written by J. Smith, the web administrator” meets the requirement. A statement like “The website manager, J. Smith, PhD, has 15 years of experience in managing websites for educational institutions and Fortune 500 corporations” does not give the expertise behind the information on the specific page containing that footer, and so does not meet the requirement.

- (4) *Show “honesty and trustworthiness.”* User comments could form a reasonable proxy for this. Since each page is different, it makes sense to provide this service for each page, in a standard footer for example (although an alternative measure could be devised for a comment service provided for the website as a whole). A formula that arises naturally is based on the fraction of pages that have a link to a user comment service that shows previous comments and allows new ones. Such a formula is:

$$100 * (\text{Fraction of pages on the website that accept and publish user comments}).$$

This formula provides a number from 0 to 100 and expresses an assessment of the honesty and trustworthiness of the website.

- (5) *Ensure ease of contacting the author.* This is achievable by having web page footers that give contact information for the person responsible for authoring the page. An email address is sufficient. A formula arises straightforwardly out of that criterion:

$$100 * (\text{Fraction of web pages on the site that provide a contact for the author}).$$

- (6) *Site is professional and appropriate.* The terms “professional” and “appropriate” are too vague to be easily measurable here, but attention to presentation quality such as lack of typos, good grammar, and quality page design provide some indication of professionalism and appropriateness. Of these, quality of page design is the hardest to measure. An inexpensively produced page can show a higher quality of design than a more elaborate design using poorly chosen graphics, so automatic measurement of page design quality is problematic. However typos such as misspellings and grammar issues are more amenable to automatic checking, especially spelling. A page could have problems in this area or not, and the proportion of pages that have no detectable problems can be a proxy that provides the basis for a measurement. A formula that would provide the desired calculation is:

$$100 * (\text{Fraction of pages on the site that have no detectable spelling or grammar errors}).$$

- (7) *Website is useful and easily used.* This is already accounted for under items 1 and 2 (Useful and Usable) above.
- (8) *Website is current.* Some pages retain currency more than others, thus needing to be updated less often. Thus, it is difficult to automatically determine whether a page is current since the date of the last update does not in itself

determine if a page is current. Pages with news may lose currency in a day, while other pages such as maps may stay current for a much longer time. However, what could be determined automatically is the date of the most recent update to the page, if the page provides that information in its footer. Using this a proxy, we can simply measure what fraction of pages provide their most recent update date in the footer. This gives the formula:

$$100 * (\text{Fraction of pages on the site that contain the date of the most recent update in their footers})$$

- (9) *Avoid promotional ads unless there is a good reason to have a few.* An academic site normally needs few ads for commercial products, and such ads are likely to detract from the quality of the site. The fraction of pages that do not contain commercial advertisements measures this characteristic, giving the formula:

$$100 * (\text{Fraction of pages on the site that contain no commercial advertising})$$

- (10) *Avoid even small errors.* This attribute focuses on seemingly minor problems like typographical errors and broken links. These errors are significant, especially when they occur more than rarely, as they “hurt a site’s credibility” [20]. Larger errors are already covered by measurement strategies listed earlier, in particular: (1) *Information accuracy should be easily verified*, (3) *Display the content expertise behind the website*, (4) *Show “honesty and trustworthiness.”* and (8) *Website is current.* In addition, grammar and spelling errors are covered under (6) *Site is professional and appropriate.* For this criterion, then, we define a measurable proxy using the fraction of links on the website that are broken. This leads to the formula

$$100 * (1 - (\text{Fraction of links on the website that are broken})) .$$

Factors (1) through (10) above all contribute to credibility, and those that are used may be averaged to determine the website’s information quality for the *Credible* dimension.

66.3.7 Valuable

This dimension is closest to the information quality dimension of *Value-Added*, which is a measure of the extent to which data is beneficial and provide advantage from their use [3]. In the context of a university website, the *Valuable* dimension implies that visit to the website should be beneficial and provide advantage for visitors. According to [13],

a non-profit site is valuable if the “user experience ... advance[s] the mission.”

Since in a higher education environment we are typically dealing with a non-profit entity, we would like to measure advancement of the mission. What is the mission? While individual institutions will have their own mission statements, generically these missions will tend to be variations of promoting and providing education. So how well does a university website promote and provide education? Authorized students may access learning materials that are hidden from the average web user, for example because they are provided by a Learning Management System that is password protected. However, learning materials available to the public would often be accessible on the web to anyone using a web browser. Such publicly accessible learning materials could truly be said to both promote and provide education. They promote it by reaching out to the general public and not just paying students, and they provide it by containing educational content. Such content might be anything from extension service articles, to course materials that individual instructors make available on the web, to scholarly works that faculty have written and make available for download from the university website.

To measure the extent to which such learning materials are provided on the website, some heuristic approach to quantification is needed. It cannot be an exact measure of total educational content because that is probably impossible to quantify reliably. It may be possible, however, to measure the fraction of university faculty involved in producing these offerings, by sampling individual faculty web pages and checking for learning materials. The fraction with links to publicly accessible educational materials is then the basis for a metric, and this metric is heuristically speaking a reasonable measure of breadth (though not depth) of offerings relative to institution size. This leads to the following formula:

$$100 \frac{(\text{number of faculty with qualifying links})}{\text{total number of faculty}}$$

66.4 Combining Information Quality Measures

66.4.1 Combining Dimensions

We have elucidated seven dimensions of data quality for a university website. These dimensions characterize the quality of a website in significantly different ways. As a result, it is possible for a site could rate highly on one dimension but poorly on another. The existence of an important dimension on which a site rates poorly suggests the site has a serious problem. Even if it rates highly on all other dimensions,

a low rating, such as on the currency or on the valuable dimension, or any other important dimension, suggests a critical weakening of the overall quality. A high quality site is required to rate highly on all important dimensions, not just almost all. Therefore taking the familiar arithmetic mean of the ratings on the different dimensions would not be the best way to combine the ratings.

Since taking the arithmetic mean is not a good way to combine the dimensions, what is a better way? The *geometric mean* is another kind of average, and it does meet the requirement that one low rating makes a big difference, usually considerably bigger than for the common (arithmetic) mean. To calculate the geometric mean, multiply the applicable n values and then take the n th root of that product. For example, consider seven dimensions all with quality ratings of 100. The geometric mean is $(100*100*100*100*100*100*100)^{1/7} = (100^7)^{1/7} = 100$, the overall quality. But one low rating of 10 instead of 100 is quite influential, giving a geometric mean and overall quality estimate of $(100*100*100*100*100*100*10)^{1/7} = (10*100^6)^{1/7} = 71.97$, quite a bit lower than 100 and in fact lower than if all seven dimensions had very questionable qualities of say 75 (in which case the geometric mean would be 75). By comparison, the common arithmetic mean would be 87.14, higher than 71.97 and not as plausible as an overall quality when a critical dimension is very low. Thus, as a method of combining the dimensions the geometric mean seems to provide a better approach than the more common arithmetic mean.

Still, a very low quality on an important dimension perhaps ought to be required to be harder to compensate for than that: 71.97 still seems high when a critical dimension is as low as 10 out of 100. Is there an even better average? The *harmonic mean* does seem to be better in this way. The harmonic mean when all seven dimensions have the value 100 is

$$7/(1/100 + 1/100 + 1/100 + 1/100 + 1/100 + 1/100 + 1/100) = 7/(7/100) = 7 * 100/7 = 100.$$

When one of the seven dimensions has the low value of 10, the harmonic mean is

$$7/(1/100 + 1/100 + 1/100 + 1/100 + 1/100 + 1/100 + 1/10) = 7/(6/100 + 1/10) = 43.75.$$

This seems more reasonable, in that a very low dimension has an overwhelming effect, yet several very high dimensions do manage to help significantly. In this, the harmonic mean differs from taking the minimum rating across the dimensions, thereby ignoring the inputs provided by the other dimensions, which does not seem desirable. We conclude that the harmonic mean models the problem the best of these three

averages. However, a more comprehensive approach to the question of how to combine dimension ratings would be a useful and important topic for future researchers.

66.4.2 Combining Multiple Measurements of a Single Dimension

For some of the dimensions we have given multiple measures that address different indicators of the quality of the dimension. These indicators are often more fungible than the dimensions themselves, in that scoring high on one sub-dimension compensates to a significant degree for scoring lower on another. If this fungibility does not hold, it suggests that the dimension is actually more than one dimension, and the indicators in question are for different dimensions. With these considerations in mind, combining subdimensions can be done using the familiar arithmetic mean.

66.5 Conclusions

University websites share many key characteristics. They typically share considerable similarities in both form and function. These sites serve as an institution of higher education's public face as well as a portal for its prospective and current students, faculty, and other staff for numerous learning and business functions. Thus, the website is a key strategic component of today's universities. Consequently, the quality of these websites is a critical factor in the business of being a university in the twenty-first century. Because of the value of high website quality to universities, it is important for universities and their Chief Information and Chief Data Officers to be able to assess their website's information quality across the relevant dimensions, calculate an overall summary quality rating, and understand specific dimensions and website characteristics needing quality improvement.

We have described a method for assessing the information quality of university websites designed to be (i) suitable given the distinctive characteristics of university websites, and (ii) readily, objectively, and quantitatively determinable. Its suitability is in part because it incorporates various relevant dimensions that both capture different aspects of quality and collectively provide reasonable coverage of the functions that go into quality of a university website. Regarding the problem of determining the quality of a site and of its component dimensions, we provide actionable methods for calculation, both of individual dimensions of overall site quality.

Determining overall site quality requires a way to combine the qualities along the various different dimensions in a meaningful way. When all dimensions are important and a low score on any one of them seriously impacts the information quality of the site, a combination method is needed

that models this. We found that the harmonic mean of the qualities of the various dimensions worked better in this way than the commonly encountered arithmetic mean, as well as better than the geometric mean.

Assessing the quality along a given dimension can often benefit from combining assessments of different aspects or indicators of that dimension. Here, typically the common arithmetic mean is a reasonable way to combine these aspects because they tend to have relatively equivalent effects on the quality along a dimension, whether positively or negatively, and whether their values are high or low. Thus the website quality assessment design we offer uses the arithmetic mean for combining different measures of the same dimension.

Universities and consultants will find the method described herein readily applicable. In addition, university websites share many common goals with websites of other kinds of organizations. Therefore, the approach we describe is expected to be readily adjusted to assessing websites of other kinds of organizations. Software could partially automate the assessment process in many though not all dimensions because many dimensions and aspects of dimensions can be measured partially or fully automatically. Web maintenance teams can use the method we describe as a guide to improving and maintaining high quality websites.

References

1. R. Wang, A product perspective on total data quality management. *Commun. ACM* **41**(2) (1998)
2. Y. Yiliyasi, D. Berleant, World oil reserves data information quality assessment analysis. In *The Proceedings of the 16th International Conference on Information Quality*, Adelaide, Australia (2011)
3. C. Fisher, E. Lauria, S. Chengalur-Smith, *Introduction to Information Quality* (AuthorHouse, 2011), Bloomington, IN
4. Information quality guidelines, online, www.eia.gov/about/information_quality_guidelines.php, accessed 13 April 2019
5. Health and human services web standards, online, webstandards.hhs.gov/standards, accessed 2019
6. B. Stvilia, L. Gasser, M.B. Twidale, L.C. Smith, A framework for information quality assessment. *JASIST* **58**(12), 1720–1733 (2007)
7. W.A. Robbins, K.R. Austin, Disclosure quality in governmental financial reports: an assessment of the appropriateness of a compound measure. *J. Account. Res.* **24**, 412–421 (1986)
8. A. Quarati, R. Albertoni, M. De Martino, Overall quality assessment of SKOS thesauri: an AHP-based approach. *J. Inf. Sci.* **43**(6), 816–834 (2016)
9. D. Tao, C. LeRouge, K.J. Smith, G. De Leo, Defining information quality into health websites: a conceptual framework of health website information quality for educated young adults, online, doi.org/10.2196/humanfactors.6455, accessed 24 April 2020
10. H.R. 5759 – 21st Century Integrated Digital Experience Act, 2018, www.congress.gov/bill/115th-congress/house-bill/5759/text, accessed 2020
11. Web content accessibility guidelines (WCAG) overview, online, www.w3.org/WAI/standards-guidelines/wcag, accessed 25 Oct. 2019
12. What & why of usability: user experience basics, online, www.usability.gov/what-and-why/user-experience.html, accessed 7 June 2019
13. Semantic Studios, User experience design, online, semanticstudios.com/user_experience_design, accessed 6 June 2019
14. A. Mentis, A.H. Turan, Assessing the usability of university websites: an empirical study on Namik Kemal University. *Turk. Online J. Educ. Technol.* **11**(3), 61–69 (2012)
15. A. Usher, Errors: what they are, and how to deal with them, online, www.physics.rutgers.edu/grad/506/errors.pdf, accessed 13 April 2019
16. Y. Yiliyasi, *The ΣIQ Method: An Information Quality Perspective on Oil Data*, dissertation, Lambert Academic Publishing, 2012
17. Y. Lee, M. Strong, B.K. Khan, R.Y. Wang, AIMQ: a methodology for information quality assessment. *Inf. Manag.* **40**, 133–146 (2002)
18. Bureau of Internet Accessibility, online, www.boia.org/products/free-wcag-2-0-aa-report, accessed 25 October 2019
19. What are the four major categories of accessibility? online, www.boia.org/blog/what-are-the-four-major-categories-of-accessibility, accessed 25 Oct. 2019
20. Stanford Persuasive Technology Lab, Stanford university web credibility research, 2004, online, <http://credibility.stanford.edu/guidelines/index.html>
21. Website navigation: tips, examples and best practices, online, www.crazyegg.com/blog/website-navigation, accessed 25 Oct. 2019

Breno Lisi Romano and Adilson Marques da Cunha

Abstract

This paper presents the Agile and Collaborative Model-Driven Development method for Web applications named WebAC-MDD. It was conceptualized to transform agile models into Web application source-codes, using a novel Unified Modeling Language (UML) profile named Web Agile Modeling Language (Web-AML). It intends to represent a proposed solution for existing and inherent problems, regarding productivity in the development of Web applications and efforts for modeling and documentation, which do not add any value to clients. The main contributions of this paper are the WebAC-MDD Method, the Web-AML profile, and the proposed automatic generation of products, from the agile models, that are of value to stakeholders.

Keywords

Code generation · MDD · UML profile · Agile · Model transformation · Web application · AMDD · AM · WebML · Modeling language

67.1 Introduction

Modeling is an important activity in every software development project. It allows thinking about problems' complexity

B. L. Romano (✉)
Federal Institute of Education, Science, and Technology of São Paulo,
São João da Boa Vista, São Paulo, Brazil
e-mail: blromano@ifsp.edu.br

A. M. da Cunha
Brazilian Aeronautics Institute of Technology, São José dos Campos,
São Paulo, Brazil
e-mail: cunha@ita.br

even before addressing them into source-codes. It holds true for different kinds of software projects like agile, prescriptive process, embedded, or any other.

Unfortunately, many efforts put into the software modeling activity have been proven dysfunctional [1]. From one side, there are many projects that do not have any kind of modeling, whether because the development team does not have the knowledge to perform such task or because they consider modeling a worthless activity from the business point of view. On the other hand, there are projects with too much modeling and unnecessary details, causing redundant or even disposable documentation.

So, it is important to deal with these difficulties and while Agile Modeling (AM) method can be used due to it defines values, principles, and practices that describe how to simplify modeling and documentation activities as part of development teams' daily routine [1].

As highlighted by Ambler (2007), what makes AM a catalyst to improve software documentation and modeling are not the modeling techniques, but how to apply them to improve productivity [1].

Also aiming to increase productivity, the Agile Model Driven Development (AMDD) was created as a approach of agile software development using extensive models before the source-code is written [2].

Additionally, according to Matinnejad (2011), AMDD can be better described as a smart compromise. Due to contrasting points of view of both agile development and MDD, an effective compromise must be reached, in order to leverage advantages of modeling without its disadvantages. This is the main goal of AMDD [3].

In this sense, the development of web applications needs more agile software development methodologies with short iterative cycles – because of short deadlines. We can then use the AMDD approach to leverage both techniques based on models and agility. This approach implies shorter iterations,

encourages stakeholders' participation and allows requisites to evolve along the project [4].

Moreover, there are many cases where the MDD is used to define transformation mechanisms to allow the generation of source-code in a specific programming language from high-level abstraction models [5].

To perform the generation of source-code from models it can be used the Unified Modeling Language (UML) because of its popularity between existing modeling language [6]. Additionally, UML define visual language with an appropriate semantics to create software structures artifacts [7]. Web Modeling Language (WebML) has introduced visual notations to conceive web applications, considering the required features for approaches based on user interfaces (front-end). These approaches represent an emerging development paradigm that focuses on final users' characteristics, in order to develop and maintain web applications [4].

In this context, it is necessary to look for solutions that allow application of the agile modeling and AMDD to design web applications. Additionally, it is important to improve productivity by generating source-codes produced models.

Therefore, the goal of this paper is to present the Agile and Collaborative Model-Driven Development Method for Web Applications (Web-AMDD Method) to transform agile models into web applications source-code using a new UML profile named Agile Modeling Language for Web Applications (Web-AML). We propose a possible solution to the following problems: existing and inherent difficulties regarding the web application development productivity; difficulties in the requirements gathering stage; and efforts put into modeling and documentation, which do not have any value to clients.

67.2 The Web Application MDD Approach

To provide the scientific and technological community with an agile collaborative model driven development for web applications, it was proposed in this paper the Web Agile and Collaborative Model-Driven Development method named WebAC-MDD shown in Fig. 67.1. The goals of each of these steps are described in ROMANO (2019) [8].

In each step of the WebAC-MDD Method presented in the previous section, visual notations are also detailed. These notations are part of the UML profile proposed in this paper named Web Agile Modeling Language (Web-AML), which depends on the needs of WebAC-MDD to make it operational. The scope of this paper is a transformation proposed to automatically generate source-code from the agile models created with the help of the WebAC-MDD method. In this sense, the Web-AML profile metamodel for the elements that represent CHIs functionalities (Step 8) and for modeling functional tests (Step 9) of the web application can be seen in ROMANO (2019) [8].

67.3 The Proposed Transformations to Source-Code Generation

From Step 10 of the WebAC-MDD method onwards, every agile model is already done. Step 10 then proposes the automatic generation of products, from agile models, that are of value to stakeholders. The goal of this is to increase the development productivity of the web application. This step justifies the use of the MDD approach in the conception of the WebAC-MDD method, as explained in Sect. 67.1.

As it can be seen in Fig. 67.2, the process to automatically generate the source-code starts with web application agile models that used the stereotype proposed by the WebAC-MDD method and included in the Web-AML profile. It is worth mentioning that these agile models must be created using a CASE (Computer-Aided Software Engineering) tool chosen by system analysts.

For this reason, regardless of the CASE tool, the WebAC-MDD method assumes that agile models are stored in files with the ".uml" extension. These files store the following: the modeled elements and their properties; the information about the elements position in the diagrams; and the additional information created by the chosen CASE tool.

Besides that, in Step 10 every agile model, regardless of the development sprint, is taken into account to automatically generate the source-code. This is done in order to assure the integration of new models with the existing ones.

The first transformation proposed in Step 10 is highlighted in Fig. 67.2 as "**Transformation 1**". In this transformation, the ".uml" files are converted to clean XML (eXtensible Markup Language) files containing only information about elements of the web application modeling. This way, the following XML files are created:

1. A file containing the domain classes model, created in Step 6 with the <Entity> stereotype, that represents entities to be stored in the database;
2. A file uniting the tests context of Step 9 created with the <TestContext> stereotype with the test cases created with the stereotype <TestCase>;
3. A file for each one of the CHIs, modeled in Step 8, using the <Page> stereotype, in order to group all modeling elements shown in Fig. 67.2 and present them in the current CHI;
4. A file containing the domain classes model, created in Step 6 with the <Entity> stereotype, that represents the entities to be stored in the database;
5. A file uniting the tests context of Step 9 created with the <TestContext> stereotype with the test cases created with the stereotype <TestCase>; and
6. A file for each one of the CHIs, modeled in Step 8 using the <Page> stereotype, in order to group all modeling elements and present them in the current CHI.

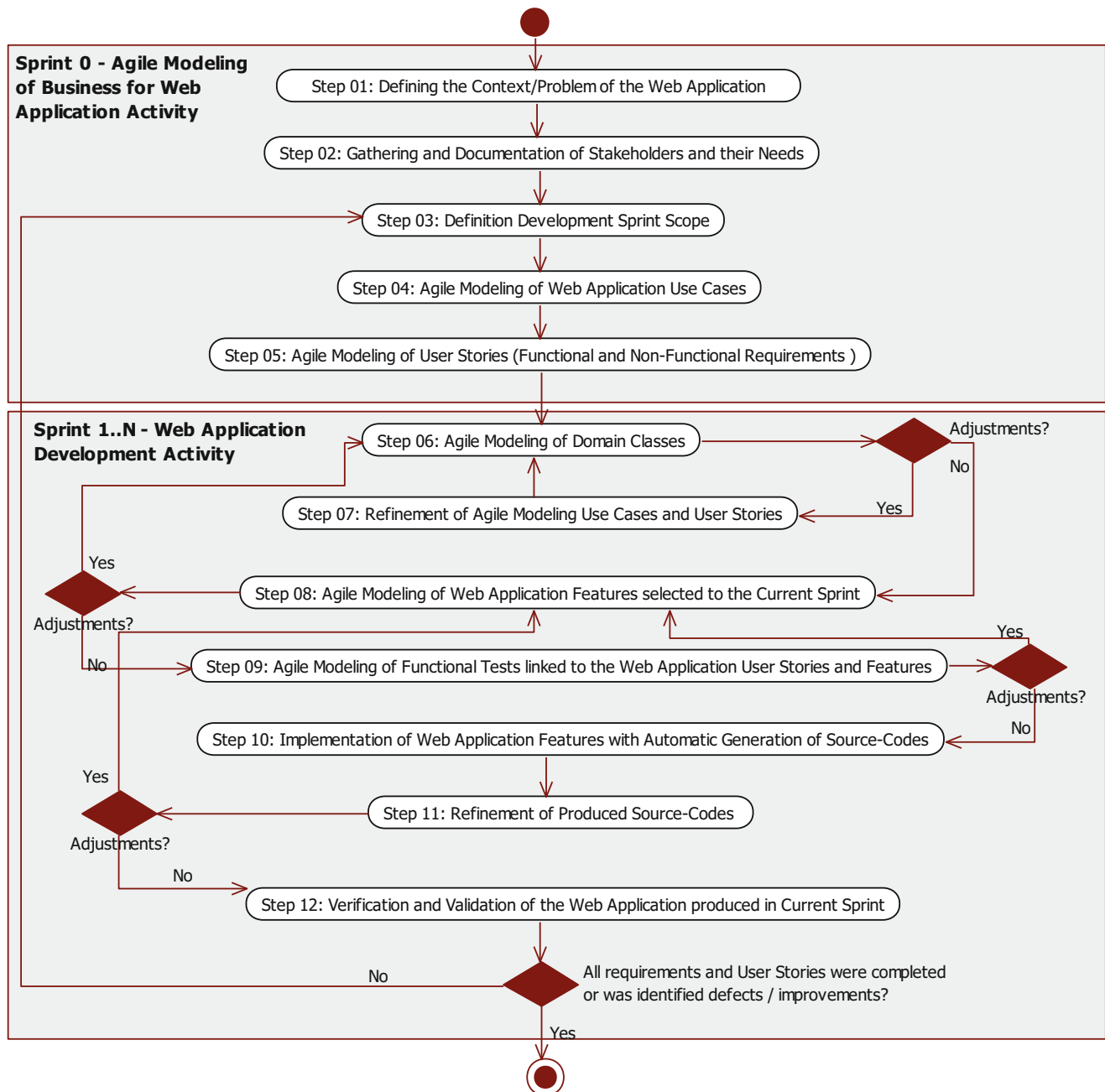


Fig. 67.1 The Web Agile and collaborative model-driven development method named WebAC-MDD

Some XML files created after “**Transformation 1**” are shown in Figs. 67.3, 67.4, and 67.5. The code shown in Fig. 67.3 illustrates the file with the web application domain classes, with details on how different classes are organized in the XML file.

The structure of the file containing all test contexts, created in Step 9, is shown in Fig. 67.4. The structure was also a result of the first transformation.

To represent each of CHIs modeled in Step 8 of the WebAC-MDD Method, we must create a XML file following the same structure shown in Fig. 67.5. It is worth mentioning

that on each page there may exist three distinct lines: valued objects, model elements, and associations between other pages. Moreover, the properties of a page and the association (links) structures are shown in the code presented in Fig. 67.5.

Once “**Transformation 1**” is done, the outputs are all agile models represented in textual files in XML format. We opted to represent the agile models in XML for the sake of interoperability of models generated by the WebAC-MDD method, facilitating data exchange between other CASE tools.

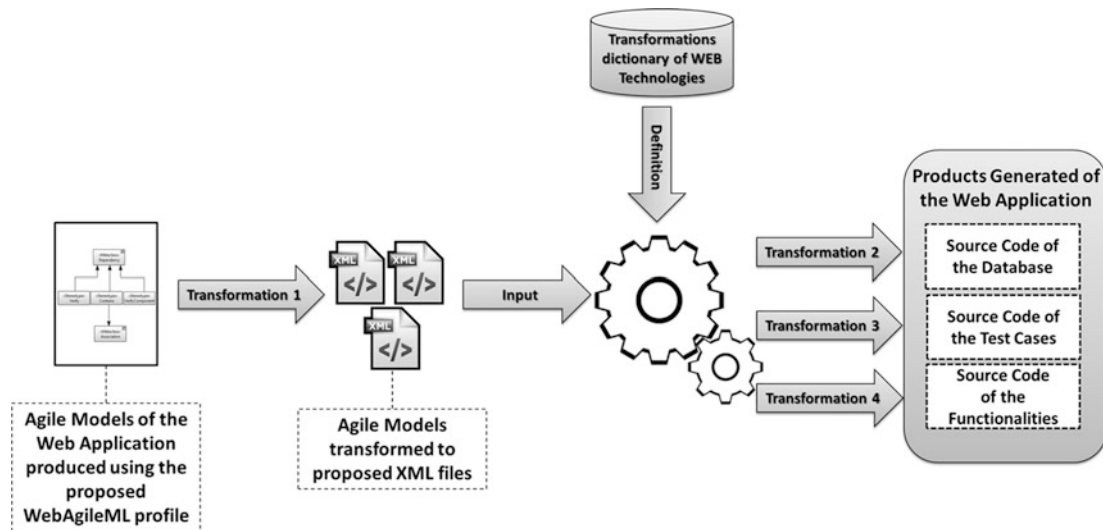


Fig. 67.2 The transformations proposed of the WebAC-MDD method to source code generation

Fig. 67.3 The XML file with web application domain classes structure

```

1 <DOMAIN name='DIAGRAM NAME'>
2 <ENTITY id='BUSINESS CLASS NAME'>
3 <ATTRIBUTE id='PROPERTY NAME 1' type='Boolean || Integer || Real || String
  '"/>
4 <...ATTRIBUTES LIST...>
5 <RELATIONSHIP id='CLASS_NAME_FROM2CLASS_NOME_TO' to='CLASS_NOME_TO'
  minCard='VALUE' maxCard='VALUE' />
6 <...RELASHIOSHIPS LIST...>
7 </ENTITY>
8 <ENTITY>
9 <...ATTRIBUTES LIST...>
10 <...RELASHIOSHIPS LIST...>
11 </ENTITY>
12 <...ENTITIES LIST...>
13 </DOMAIN>

```

To do the next transformations defined in Fig. 67.2, we need to choose some web technologies. We also need that transformations for such technologies must exist in the transformation dictionary. This is required, in order to generate the web application source-code.

In the transformation dictionary, for each web technology there must exist a mapping between every model element generated after “**Transformation 1**” with its textual representation in the chosen web technology. From this mapping, we can automatically generate the web application source code.

“**Transformation 2**” uses the XML file with the domain class agile models shown in Fig. 67.3 and the following steps to automatically generate SQL source-code (the goal in this case is to create the database structure, with its tables, attributes, and relationships):

1. Convert the domain class XML file into a single SQL file;
2. Read the whole domain class XML file and do the following:

- (a) Generate the structure of each table with the SQL domain class name, represented by the “ENTITY” separator;
 - (b) Generate an identifier attribute with an integer and self-incrementing data type. The name of this attribute must be “CLASS_NAME_ID” as tables’ primary keys;
 - (c) Generate each one of the attributes in the domain class. The “ATTRIBUTE” separator represents them, with a name defined in the “name” property and data type defined in the “type” property. They must be converted to a corresponding SQL type; and
3. Read the domain class XML file again from the beginning, to create every existing relationship between classes. They are represented by the “RELATIONSHIP” separator and are considered as foreign keys. If the relationship is “1-N” or “0-N”, we must create a foreign key in the table with N multiplicity with the same name of the primary key of the table with multiplicity 1. If an “N-N” relationship does exist, we must create a third table concatenating the name

Fig. 67.4 The XML file with the web application tests structure

```

1 <TEST name='DIAGRAM NAME'>
2 <TESTCONTEXT name='TEST CONTEXT NAME' description='TEST CONTEXT
  DESCRIPTION'>
3 <USERSTORY name='USER STORY TO BE VERIFY' />
4 <...USER STORIES LIST...>
5 <TESTCASE name='TEST CASE NAME' description='TEST CASE DESCRIPTION'
  assert='ADOPTED ASSERT' method='MEIHOD TO BE USED'
  verifiedComponent='COMPONENT TO BE VERIFIED'>
6 <TESTINPUT name='INPUT PARAMETER NAME' type='Boolean || Integer || Real
  || String' />
7 <...LISTA DE PARÂMETROS DE ENTRADA...>
8 <TESTOUTPUT name='OUTPUT PARAMETER NAME' type='Boolean || Integer ||
  Real || String' />
9 </TESTCASE>
10 <...TEST CASES LIST...>
11 </TESTCONTEXT>
12 <TESTCONTEXT>
13 <...USER STORIES LIST...>
14 <...TEST CASES LIST...>
15 </TESTCONTEXT>
16 <...TEST CONTEXTS LIST...>
17 </TEST>

```

Fig. 67.5 The XML file with the structure of a web application CHI page

```

1 <PAGE name='PAGE NAME' homepage='YES||NO' landmark='YES||NO' limited
  ='YES||NO'>
2 <...VALUE OBJECTS USED ON THE PAGE...>
3 <...MODELS ELEMENTS USED ON THE LIST...>
4 <LINK id='AUTO INCREMENT NUMBER' to='PAGE NAME' automatic='YES||NO'>
5 <PARAMETER name='PARAMETER NAME' type='Boolean || Integer || Real || String
  || ValueObject' />
6 <...PARAMETERS LIST OR VALUE OBJECTS LIST...>
7 </LINK>
8 <...LINKS PRESENT ON THE PAGE...>
9 </PAGE>

```

of the two tables following the “ENTITYA2ENTITYB” standard. This third table must have as foreign keys the primary keys of the two other tables. Another integer and self-incrementing type must be created. Its name must be “ENTITYA2ENTITYB_ID” and it must be the primary key of the third table.

“**Transformation 3**” uses the XML file with all test contexts and their respective test cases, in order to automatically generate the web application source-code. The steps required to do this transformation are the following:

1. Convert the XML file with the test contexts to a test source-code file. A test source code must be created for each context test identified by the “TESTCONTEXT” separator;
2. For each test source-code of a test context, do the following:
 - (a) Generate the structure of a class made of a method in the test case, which name in the test context is the class name and the test case names are the class methods.

The test cases are identified by the “TESTCASE” separator;

- (b) Generate the behavior of each test case method by doing the following:
 - (i) Instantiate the input parameters of the test case method using the “TESTINPUT” separator. The name and data type must be taken into account, and the latter must be converted to an appropriate type depending on which technology is being used;
 - (ii) Instantiate the parameter of the expected result using the “TESTOUTPUT” separator. The name and data type must be taken into account, and the latter must be converted to an appropriate type, depending on which technology is being used;
 - (iii) Instantiate the parameter that represents the output of the method defined in the test case by the “method” property; and
 - (iv) Apply the assert defined in the test case by the “assert” property and compare the result with

what was expected. If the results are equal, the test completion must return true. Otherwise, it must return false.

Finally, “**Transformation 4**” takes into account every XML file that textually represents the web application pages that include the XML codes of each model element. Then a number of steps must be followed to automatically generate the functionalities source-code:

1. Convert each of the XML files that represent the web application CHIs, defined by the “PAGE” separator, to three distinct files, using the Model-View-Controller (MVC) standard:
 - (a) View: HTML5 source-code to show data to the final users;
 - (b) Model: source-code with all CHI logic, method behaviors, and database interactions; and
 - (c) Controller: source-code that connects the pages viewed by the users (View) to their logic (Model);
2. For each XML file that represents a page, we must:
 - (a) Generate, in the view file, the source-code of classes that represents each of the objects identified by the “VO” separator. The class-name must be the same as in the “name” property and the class-attributes must also be considered. It is identified by the “ATTRIBUTE” separator. Besides that, we must also generate the business classes that are directly linked to the model elements.;
 - (b) Generate, in all three files (View, Model, and Controller), source-code to represent each model element of the page. These elements are structured as mentioned in Step 10. Given this scenario, the transformation dictionary must have a mapping between the model elements and the web application source-code, which in turn must be created according to the agile models; and
 - (c) Generate source-code of the links between the web pages and/or the model elements. The “LINK” separator must be used to create the links and the “to” property to identify link’s destination. If there are parameters to be sent through the link, the “PARAMETER” separator must be used in order to specify parameters’ names and data types.

It is worth mentioning that the “look & feel” aspect of the web application is not included in Step 10 of the WebAC-MDD Method. The web application developers must manually create it.

67.4 Conclusion

This paper aimed to present the Web-AMDD method to transform agile models into web application source-code, using a new modeling language named Web-AML (UML profile). We proposed an approach to deal with the difficulties regarding web applications development productivity and efforts spent from modeling and documenting – activities that do not have value to clients.

The Web-AML profile was conceived to allow the agile modeling of web applications. It takes into account agile development and Agile Modeling (AM) characteristics that were not included by UML, SysML, and WebML.

Since the Web-AML is a new UML profile, it can be used with any CASE tool, in order to allow the interoperability of models created with the WebAC-MDD method. This way, the data exchange between tools is facilitated. Moreover, both the WebAC-MDD method and the Web-AML profile can be used in a stand-alone manner.

As future work, we aim to use the WebAC-MDD method and the Web-AML profile in a real case study and other projects to receive feedback from the people involved in such projects. Finally, we believe that an optimization of the WebAC-MDD method is possible, in order to create models used only for the generation of the web application source-code. In this sense, we also aim to adjust the steps proposed in the method after being used in other projects. It is important to notice that this paper shows that using the WebAC-MDD method is a viable alternative to increase the web application development productivity and decrease efforts put into modeling and documentation.

References

1. S.W. Ambler, Agile model driven development (amdd). Xootic Magazine (2007)
2. S.W. Ambler. Agile Software Development at Scale. In: Meyer B., Nawrocki J.R., Walter B. (eds) Balancing Agility and Formalism in Software Engineering. CEE-SET 2007. Lecture Notes in Computer Science., Springer, Berlin, Heidelberg. 5082 (2008) https://doi.org/10.1007/978-3-540-85279-7_1
3. R. Matinejad, Agile model driven development: an intelligent compromise. In: *IEEE. Software Engineering Research, Management and Applications (SERA), 2011 9th International Conference on. [S.L.]*, 2011. p. 197–202
4. X. Liang, I. Marmaridis, A. Ginige. Facilitating agile model driven development and end-user development for evolving web-based workflow applications. In: *e-Business Engineering, 2007. ICEBE 2007. IEEE International Conference on. [S.l.: s.n.]*, 2007, pp. 231–238
5. B.L. Romano; G.B.E. Silva, A.M.D. Cunha, W.I. Mourao, Applying MDA development approach to a hydrological

- project. In: *IEEE. Information Technology: New Generations (ITNG), 2010 Seventh International Conference on. [S.L.]*, 2010, pp. 1127–1132
6. B. Rumpe, Agile test-based modeling. *International Conference on Software Engineering Research and Practice (SERP06)*, v. 1, pp. 10–15 (2014)
 7. Object Management Group. OMG unified modeling language v2.5. jun. 2015. Available at: <http://www.uml.org/>. Accessed in: 29 mar. 2016
 8. B. L. Romano, A. M. da Cunha. A framework for web applications using an agile and collaborative model driven development (AC-MDD). *Acta Scientiarum. Technology*, 41(1), e38349 (2019) <https://doi.org/10.4025/actascitechnol.v41i1.38349>

Index

- A**
- AAC-devices
 - Android and AAC applications, 171
 - application logic, 174–175
 - ASD prevalence, 171
 - backend text to speech functionality, 174
 - computational limitations, 175–176
 - for computing device, 172
 - cost and ease of assembly, 175
 - design process, 172
 - device customization, 175
 - GUI design, 174
 - hardware limitations, 176
 - high-tech devices, 170–171
 - iPad and AAC applications, 171
 - low tech devices, 170, 171
 - OS installation and hardware assembly, 174
 - power delivery system, 172
 - software choices
 - operating system, 173
 - programming language, 173–174
 - unaided devices, 170
 - user interface system, 172–173
 - Abstract Intent Test (AIT), 223
 - Access control contracts (ACC), 416
 - Access control models
 - ABAC, 416
 - CapBAC, 416
 - centralized architecture, 416
 - RBAC, 416
 - Accessibility IQ, 513
 - ACM Digital Library, 12
 - Active shape models (ASMs), 464
 - Actor Critic, 13
 - Adaptive learning systems, 356
 - Adjusted Rand index (ARI), 405, 406
 - Adopting monitoring systems, 398
 - Advanced encryption standard (AES), 121, 122, 422, 423
 - Advanced persistent threats (APT), 116, 117
 - AES cryptosystem, 422
 - Agile
 - development, 215
 - modeling (AM), 519, 520, 524
 - software modeling activity, 519
 - Agile Model Driven Development (AMDD), 519–520
 - Agile Software Development and Design Thinking, 399
 - Agriculture
 - ML, 34
 - origin, 33
 - prediction systems, 35
 - smart agriculture, 33–34
 - sustainability, 34
 - Ahead of Time (AOT), 229
 - AI-driven Blockchain, 433
 - Aircraft, 20, 385, 386
 - Alerting devices, 169, 170
 - AlexNet model, 452
 - Alternating least squares (ALS), 490–492, 495
 - Amazon’s review data, 490
 - Ambient assisted living (AAL), 163, 164, 166
 - Ambient intelligence (AmI)
 - AAL technologies, 163, 164, 166
 - intelligent computing, 163
 - MS protocol, 163–166
 - PVI, 163, 165, 166
 - Analytical techniques, 3
 - Analytic hierarchy process (AHP), 77, 510
 - The Analytic Hierarchy Process (AHP), 198
 - Android devices, 171–172
 - Android smartphone application, 223
 - Android testing, 274
 - Anti-trafficking organizations, 498
 - Antivaccines, 31
 - Apache Spark, 490
 - Application Programming Interface (API), 401, 483
 - Application virtual machine, 401
 - AQUACROP simulations, 35
 - Aqua-Sim, 96
 - Arbitrary structured neural networks (ASNN), 324
 - Army Integrated Telematics Center (CITEx), 200
 - Artificial intelligence (AI)
 - on cloud services, 39
 - COVID-19, 28
 - Decision Trees, 16
 - definition, 40
 - drone racing, 20
 - economic growth rates, 40
 - IDSS, 3, 4 (*see also* Intelligent decision support systems (IDSS))
 - IoT-based problems, 430
 - IoT-driven organizations, 430
 - machine learning, 432, 433
 - market analysis, 39–40
 - and ML models, 11
 - popularity, 39–40
 - smart agriculture, 34
 - smart city, 433
 - SMR, 12
 - solutions, 39
 - vehicle flow, 11
 - vehicle traffic control, 12

- Artificial neural networks (ANNs), 13, 323
 - cognitive profiles, 356, 358
 - data mining systems, 357
 - Feed Forward Networks, 355
 - forecasting systems, 358
 - Hamming Network, 355
 - Hopfield Network, 355
 - Kohonen Network, 355
 - and mathematical learning models, 357
 - natural biological networks, neurons, 355
 - in pedagogy, 356
 - personal characteristics, 357
 - Recurrent Networks, 355
 - scheme, 357
 - training, 355–356
 - working, 355
- Assistive listening devices (ALD's), 169
- Attacks
 - on cloud-based web services, 117
 - DoS attack, 116
 - on SCADA system, 117
 - in security assessment taxonomy, 117
 - on VLANs, 115–118
- Attribute based access control (ABAC), 416
- Audio analytics, 485
- Auditing, 423–425
- Augmentative and Alternative Communication Devices (AAC's), 169
 - See also* AAC-devices
- Authenticated Data Structures (ADS), 109–110
- Autism Spectrum Disorder (ASD), 171
- Auto backup, 423
- Automated software tests, 70
- Automatic learning algorithms, 37
- Automatic learning techniques, 35
- Automatic Meter Reading Technology (AMR)
 - automatic collection, consumption and diagnostic data, 436
 - blockchain (*see* Blockchain based AMR)
 - consumer friend, 435
 - cryptographic technologies, 436
 - database, 435
 - data transfer, 436
 - ethereum, 436
 - linear error-correcting block codes, 436
 - NB-IoT, 436
 - online applications, 435
 - peer-to-peer topology, 436
 - tampering/malfunction, 436
 - WIFI, 436
 - wired and wireless systems, 436
- Automatic Quality Engineering of Event-driven Software (Auto-QUEST), 284, 286–287
- Autonomous supply chain, 430
- Autonomous systems, 20
- Autonomous Underwater Vehicles (AUVs), 94
- Autonomous vehicles, 222, 226, 293
- Aviation, *see* Drones

- B**
- Bag-of-words technique, 459
- Basic closed-loop matrix, 256, 257
- Bayesian belief network (BBN), 137–143
 - CVSS BBN, 80, 81
 - interactive model, 80
 - NVD data distribution and CVSS, 79
 - risk estimation, 77
 - sensitivity, 82
 - topology, 79, 81, 82
- Big data, 428, 489, 490
 - application domain, 479
 - applications, 433
 - Bigs, 482
 - businesses, 432
 - definition, 480
 - semi-structured data, 480
 - social media-based platforms, 480
 - structured data, 480
 - “Sunflower Model of Big Data,” 480, 481
 - “3 Vs,” 480
 - unstructured data, 480
- Big data analytics (BDA)
 - audio analytics, 485
 - big social data, 479, 482
 - categories, 484
 - data science, 480
 - diagnostic analytics, 485
 - image analytics, 485
 - monitoring, social networking websites, 480
 - in social media, 480, 484–485
- Big social data, 479, 481, 483–484
- Binary classification, 473
- Binghamton University 3D Facial Expression (BU-3DFE), 465
- Bioinformatics, 307
- Bioinformatics data, 179–183
- Bioinformatics projects, 180–183
- Biological processes, 307
- Biometrics, 469
- Biosensor data, 398
- Bitcoin, 498
 - attacks, 422
 - decentralized digital cryptocurrency, 421
 - government-based banking systems, 421
 - physical money, 421
 - private key, 421, 422
 - public key, 421, 422
 - transactions, 421
- Bitcoinj java package, 424
- Bitcoin Key Manager (BKM), 423, 424
- Bitcoin Two-Factor Authenticator (B2FA), 423
- Bitcoin Wallet Auditor (BWA), 423, 424
- Bitcoin Wallet Backuper (BWB), 423, 424
- Bitcoin Wallet Encryptor (BWE), 423, 424
- Bitcoin wallets
 - accountability, 425
 - attack, 422
 - availability, 422, 425
 - confidentiality, 422, 425, 426
- Bitcoin wallet security system (BWSS)
 - accountability, 425, 426
 - architecture, 423
 - availability, 422, 425, 426
 - backup, 423
 - B2FA, 423
 - BKM, 423, 424
 - BWA, 423, 424
 - BWB, 423, 424
 - BWE, 423, 424
 - BWS, 423
 - confidentiality, 422, 425, 426
 - encryption, 422, 425
 - flow chart, 424
 - goal, 422

- integrity, 422
 - Java programming language, 426
 - password-based solution, 422, 423
 - SAL, 424
 - security and privacy challenges
 - Bitcoin Wallet security, 422
 - BlueWallet, 422
 - damage, virus, 422
 - double spending, 422
 - mining attack, 422
 - network attack, 422
 - stolen and use, hacker, 422
 - wallet damage, 422
 - system implementation, 424
 - Trust Zone technology, 422
 - two-factor authentication, 422
 - BitcoinWallet Store (BWS), 423
 - BLAST algorithm, 311
 - Blockchain
 - applications, 433
 - blocks, 436
 - challenges, 429
 - chronological articles list, 415
 - cloud computing, 429, 431
 - decentralized network, 436
 - digital signature, 430
 - features, 435
 - inclusion and exclusion criteria, 414
 - information exchange, 428
 - integration, 431
 - IoT, 416, 429
 - limitations, 418, 419
 - Merkle trees, 108, 112
 - network, 413
 - operation and structure, 428, 430
 - permissioned blockchain, 418, 419
 - privacy, 429
 - public, 418, 419
 - public keys, 436
 - public ledger concept, 413
 - quality evaluation, 414
 - research process, 413
 - research questions, 413, 416
 - secure infrastructure, 416
 - smart city, 433
 - smart contracts (*see* Smart contracts)
 - solve scalability, 429
 - transactions, 435, 436, 439
 - transparency, 432, 436
 - transparent, 436
 - UAV⁴ network, 417
 - Blockchain based AMR
 - application server environment, 438
 - architecture, 437
 - design, 437
 - DoS attacks, 438, 439
 - ethereum smart contracts, 438
 - framework, 437
 - implementation, 439
 - latency, 438, 439
 - packet generation tools, 438
 - proposed framework, 437–438
 - response time, 438, 439
 - writing, 438
 - Block cipher, 122, 123, 125
 - BlueWallet, 422
 - BOIA, *see* Bureau of Internet Accessibility (BOIA)
 - BookClub artistic makeup data set, 452
 - Boundary approximation
 - application areas, 247
 - GIS, 247
 - polygonal approximation, 247–249
 - polygonal chains, 247
 - visibility, 249–252
 - Brazilian Data Protection General Law (LGPD), 85
 - Breast cancer
 - cancer domain, 4, 5, 7
 - ontology in Protégé 5.5, 5–7 (*see also* Protege (open-source software))
 - prediction system, 4–5
 - risk factors, 5
 - Bridges supercomputer, 111
 - Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm, 225
 - Building Information Model (BIM), 351
 - Bureau of Internet Accessibility (BOIA), 513
 - Business organizations, 482
- C**
- California Consumer Privacy Act (CCPA), 85
 - Cancerous tumors, 4
 - Capability based access control (CapBAC), 416
 - CARS, *see* Containerized Amazon Recommender System (CARS)
 - Center for Internet Security (CIS) controls, 61
 - Centers for Disease Control and Prevention (CDC), 294
 - Central Regulation Center (CEHR), COVID-19, 399
 - Chatbots, 433
 - Chief Marketing Officers (CMO), 479
 - Ciphertext, 121–125
 - Citizens engagement
 - in public administration, 209
 - in smart cities, 211
 - City planning, 348
 - Civil Service
 - criminals, 203
 - data mining, 204
 - prevention, 204
 - statistics and computing, 204
 - systematic mapping, 204
 - Classification task, 457
 - Classification techniques, 460
 - ClimWat, 36
 - Clinical bioinformatics, 180, 182, 183
 - Clinical health care, 181, 182
 - Clinical MRP, 400
 - Clinical Reasoning ontology (CRO), 4
 - Cloud computing, 385, 432
 - Clustering, 404–407, 498
 - CNN models, face recognition
 - accuracy, 453, 454
 - BookClub artistic makeup data set, 452
 - image class, 453
 - makeup, 455
 - misidentified image classes, 452, 453
 - occlusions, 455
 - precision, 454, 455
 - trusted accuracy, 455
 - unconstrained accuracy, 455
 - VGG19 model, 452, 453
 - Code generation, 520–524
 - Cognitive profiles, 356, 358
 - Collaborative filtering, 490

- Combinatorial coverage (CCov), 275
- Combined classifiers
 - AdaBoost, 461
 - Bagging, 461
 - classification techniques, 460
 - data set, 459
 - majority vote, 458
 - MBTI, 458, 461
 - Myers-Briggs theory, 459
 - oversampling techniques, 460, 461
 - pre-processing, 459, 461
 - random oversampling, 461
 - resampling, 460
 - social network, 457
 - text mining, 459
- Common attack pattern enumeration and classification (CAPECTM), 137–139
- Common vulnerability scoring system (CVSS)
 - BBN exploitability, 80
 - BBN topology, 80–81
 - for COTS software systems, 78
 - CVSS Risk Estimation Model, 78
 - CVSS V2 columns, 79
 - with CWSS factors, 139
 - information derives, 79
 - NVD data, 78
 - and NVD data distribution, 79
 - for quantifying potential security threats, 137
 - vulnerability, 138
- Common weakness risk analysis framework (CWRAF), 143
- Common weakness scoring system (CWSS), 139
- Communicating Sequential Processes, 313
- Compartmental models, 388
- Competitive ratios, 266–271
- Composite IQ, 509–511
- Compound emotion (CE), 465
- Computational biology, 307
- Computational cost, 121, 123, 128
- Computer-aided diagnosis, 463
- Computer-Aided Software Engineering (CASE) tool, 520, 521
- Computer Engineering, 378, 379
- Computer graphics, 247
- Computer science (CS), 291
- Computer vision (CV), 463, 464
 - deep learning, 469, 475
 - extracted hand-crafted features, 470
 - feature extraction, 470
 - ML, 469
 - processes, 470
- Computer worms
 - classification
 - academic database search, 52
 - behavioral analysis, 51
 - “classes,” 54, 56
 - database distribution, 53
 - EDOWA, 51, 54, 55
 - evasion, 54, 56
 - Internet, 54
 - obfuscation techniques, 50
 - taxonomies, 51–53
 - “types,” 54, 56
 - classification proposal, 50
 - Creeper worm, 49
 - history, 49, 50
 - on obfuscation patterns, 50
 - spread of worms, 50–51
 - SRP, 49
 - worm detection systems, 50
- Compute Unified Device Architecture (CUDA), 324, 332–333
- Computing correlations, 503
- Concurrent models, 258, 260
- Conda, 490–491
- Conditional random fields (CRFs)
 - Android and iOS, 221
 - Android device, 225
 - AndroidWikipedia app, 221
 - application’s execution, 222
 - context aware applications, 222
 - ContextMon, 222
 - cost effective testing, 221
 - data collection and analysis, 222
 - deep neural networks, 222–223
 - empirical study
 - data description, 224
 - experimental setup, 224–225
 - research questions, 224
 - study participants, 224
 - Google Play, 221
 - GUI testing, 223
 - and HMM, 222
 - learning, 225
 - ML algorithms, 223
 - ML techniques, 222, 225
 - real-world context, 222
 - research methodology
 - context modeling, 224
 - data collection process, 223
 - data preprocessing, 223–224
 - smartphone app, 221, 222
 - threats to validity, 226
 - transitions, 225, 226
 - web applications, 223
- Confidentiality, 422, 425, 426
- Consecutive-sequence combinatorial coverage (CSCov), 275
- Constant-time implementation, 131, 133
- Containerized Amazon Recommender System (CARS)
 - ALS, 490
 - approach, 490–491
 - big data, 489
 - Conda, 490–491
 - content-based approach, 490
 - data analysis, 491–495
 - data visualizations, 491–495
 - Docker, 489, 491
 - entertainment platforms, 490
 - environment, 489
 - implementation, 490
 - MLlib library, 490
 - prototype, 489
 - Spark the Definitive Guide*, 490
 - user-data-driven recommendations, 490
 - user preferences, 490
- Content analysis, 150, 362, 486
- Content-based approach, 490
- Continuous state machine, 266, 267, 270
- Control cloud applications, 432
- Controls, CIS, 61
- Convolutional neural networks (CNNs), 446, 464, 467, 469–471, 473–475
- Coronavirus 2, 397

- The Corona Virus Disease 2019 (COVID-19)
 - AI and I4.0 technologies, 28, 31, 385
 - algorithmic solution, 28
 - in Australia, 384
 - in Brazil, 384
 - in computed tomography, 27
 - deaths, 397
 - drone-based systems, 384
 - economics of vaccination in India, 392
 - epidemiology, 388
 - exploratory research, 27
 - humanity, 384
 - impact of vaccination, 391, 395
 - Implement Risk Management tools, 30
 - in India, 384
 - from Influenza A/B, 30
 - medical services, 75
 - outbreak, 31
 - pandemic, 70
 - person-to-person transmission, 397
 - respiratory rehabilitation, 384
 - scientific gaps, 28–30
 - social distance, 383
 - software and hardware system, 386
 - symptoms monitoring, 398–399
 - and technology, 28
 - UAV-based systems, 383–386
 - unexpected appearance, 384
 - vaccination campaign, 387, 392–393
 - Virtual Hospital (*see* Virtual Hospital)
 - in Wuhan, 27
 - Coronaviruses, 397
 - Coronavirus transmission, 27
 - Correlations, 503, 505–507
 - Counter Trafficking Data Collaborative (CTDC), 498, 500
 - COVID-19 epidemic
 - biological and sociological factors, 185
 - city organization, 187
 - disease state, 186
 - hybrid uses, 186
 - immunities, 192
 - infected individuals, 187
 - lockdown, 190, 191
 - mathematical models, 185
 - measuring success, 188
 - memory parallelism, 185
 - parallelization, 188, 192
 - parameter estimation, 188
 - simulating movement, 187
 - simulation, 186, 189
 - speedup, program, 190
 - susceptible individuals, 186
 - 2D space, 186
 - COVID-19 remote monitoring system, 402
 - COVID-19 restrictions, 22
 - Creeper worm, 49
 - Criteria classification, ML algorithm, 40–45
 - Critical infrastructure, 69
 - Crop management, 33
 - CropWaterNeed approach
 - climatic data, 36
 - climatic parameters, 36
 - CLIMWAT, 36
 - COVID-19 crisis, 36
 - CROPWAT program, 36
 - data aggregation, 36
 - experimentation, 37
 - feature engineering, 36–37
 - learning algorithms, 37
 - merit, 34
 - ML, 35
 - Scikit Learn library, 37
 - training data, 37
 - water requirements for crops, 36
 - Cross-validation, 460
 - Crowdsourcing, 216
 - Crows sensing, 432
 - Cryptocurrency, 436, 498
 - Cryptography, 123
 - algorithm/method, 127
 - classical, 127
 - Gauss-Jacques method, 128
 - Gauss-Jordan model, 129
 - Hill cipher, 128 (*see also* Hill cipher)
 - numerical methods, 128, 129
 - quantum, 127
 - symmetric/one-key, 127, 128
 - technologies, 436
 - Cryptosystem, 121, 123, 423–425
 - CudaNode, 326–327
 - Culturally responsive teaching, 369
 - Curricular innovation, 503, 506, 508
 - Customer engagement, 433
 - CVE (Common Vulnerabilities and Exposures), 70
 - CVSS Risk Estimation Model, 78, 138
 - Cyber-physical spaces, 78
 - Cybersecurity
 - analysis in nodes (*see* Nodes, cybersecurity analysis)
 - DICOM protocol (*see* DICOM (Digital Imaging and Communications in Medicine) protocol)
 - EDOWA system, 50
 - Internet, 54
 - Merkle tree (*see* Merkle tree)
 - NADTW, 51
 - PACS-DICOM servers, 70
 - risk management, 77
 - SRP, 49
 - Cyber Survivability Endorsement (CSE), 78, 138
 - Cyber Survivability Risk Category (CSRC), 78
 - Cyber Survivability Risk Posture (CSRP), 138
 - CycleGAN, 449, 450
- D**
- Dark web, 498, 501
 - Database Management System (DBMS), 401
 - Data collection
 - E-NEST remote learning, 370
 - Data distribution, 504
 - Data-driven identification, 503–508
 - Data extraction, 504
 - Data framework, 79
 - Data immutability, 414
 - Data integrity, 413, 416
 - Data preprocessing, 504
 - Data processing, 294
 - Data quality, 515
 - Data rate (DR), 92, 97–99
 - Data science, 480, 501
 - Data set, 490–494, 505–506
 - Data structure, 107–110, 112
 - Data transmission, 247

- Data visualizations
 - helpful review data, 492, 494
 - items over time, 491, 492
 - prediction performance, 492, 495
 - summary statistics, 491, 493, 494
 - Decision-making, 11, 13, 16
 - Decision support system
 - in breast cells, 3
 - Decision tree (DT), 460, 504
 - smart agriculture, 34, 37
 - DEC PDP-10 (Digital Equipment Corporation Programmed Data Processor model 10) computers, 49
 - Deep learning (DL)
 - closed-end classification problems, 475
 - CNN, 473, 475
 - disadvantage, 475
 - face recognition, 451
 - facial expression approaches, 464
 - image classification, 469, 473
 - Deep neural networks, 222–223
 - Deep Q-learning, 13
 - Denver Intensity of Spontaneous Facial Action Database (DISFA), 465
 - Department of Health and Human Services (HHS), 509
 - Dependent variables, 241
 - Description Logic (DL), 4
 - Descriptive analytics, 485
 - Detecting process, 467
 - Deterministic models, 186, 388, 391
 - DICOM (Digital Imaging and Communications in Medicine) protocol, 70–74
 - Difference-of-Gaussian (DoG), 472
 - Diffusion smooth-based methods
 - DrImpute, 404
 - kNN-smoothing, 404
 - MAGIC, 404
 - Digital health, 30
 - Digital services, 511
 - Digital Signal Processing (MLDSP), 30
 - Digital Signature Algorithm (DSA), 131
 - See also* Rainbow (signature schemes)
 - Digital technology, 369
 - Dimension IQ
 - currency, 516
 - data quality, 515
 - geometric mean, 516
 - harmonic mean, 516
 - identifying, 510–511
 - measurements, 511
 - single, 516
 - valuable, 516
 - Dimension reduction, 406
 - Discrete event systems, 255, 256
 - Discrete Global Grid Systems (DGGs), 352
 - Disease detection, 463, 466
 - Disruptive technologies
 - applications, 433
 - business models, 428
 - challenges, 428
 - cloud, 428
 - Hype Cycle, 428, 429
 - integration and convergence, 431
 - IoT, 428
 - opportunities, 432
 - social media, 428
 - Distributed applications (DApPs), 436
 - Distributed ledger, 413, 416
 - Distributed machine environments, 265
 - Distributed Power Generation (DEG) projects, 199
 - DNA cryptography, 123
 - Docker, 176, 417, 489, 491, 495
 - Docker Swarm tool, 417
 - DoG SIFT descriptors, 474
 - DoS (Denial of Service) attack, 116
 - DoS attacks, 438
 - DOT graph format file, 257
 - Double spending, 422
 - D-P algorithm, 247–248
 - DrImpute, 404, 405
 - Drone racing, 21
 - Amazon, 19
 - and 5G, 24
 - human-in-the-loop, 20
 - IEEE, 20
 - military, 20
 - small UAVs, 20
 - Drones
 - Amazon, 19
 - challenge, 19–20
 - and flight simulation, 19
 - flight simulator, 19
 - immersive experience, 19
 - NASA-made diagram, 20
 - Parrot AR 2.0 drone, 19–21
 - personal medical kit for COVID-19, 384
 - precision agriculture, 34
 - “Send Video” use case, 22
 - simulator movement, 22, 23
 - for surveillance, 384
 - UAVs, 384
 - use case diagram, 21, 22
 - WiFi, 22
 - Dynamic neural network models, 35
 - Dynamic Programming (DP), 13
- E**
- E-commerce, 207, 339, 489, 498
 - Education
 - HEIs, 373, 377, 379
 - industry, 509–510
 - professional ethics, 374–375
 - technology, 361
 - Efficient Worm Attack Detection System (EDOWA), 50
 - E-health, 157
 - EIA, *see* Energy Information Administration (EIA)
 - Electrical energy, 265
 - Electronic health interventions, 155
 - Electronic health records (EHR), 179, 180
 - Emerging technology, 361, 364
 - Empathy
 - capacity, 361
 - description, 362
 - development, 361–363
 - immersion and embodiment, 363, 364
 - IRI, 363
 - VR (*see* Virtual reality (VR))
 - Encryption, 422, 425, 426
 - End-to-end delay (E2ED), 97–102
 - Energy consumption (EC), 97–102
 - Energy Information Administration (EIA), 509
 - E-NEST remote learning
 - benefits, 370

- challenges, 369
 - conceptual framework factors, 369
 - future STEM Education projects, 371
 - internships and mentorships, 367–371
 - Noyce summer workshops, 368, 370
 - online instruction, 370, 371
 - online learning tools, 368, 369
 - online recording of meetings, 370
 - project management, 369
 - recruitment and mentorships, 370–371
 - remote teaching and learning, 370–371
 - STEM faculty, 367
 - three-tiered project model, 368
 - Zoom and Microsoft Teams apps, 368
 - Enrollment, 510
 - Ensemble classifiers, 457
 - Enterprise 2.0, 482
 - Enterprise Architecture (EA) Model, 200
 - Entertainment platforms, 490
 - ε -Rectangles, 249, 251–252
 - Ether, 436
 - Ethereum, 436, 437
 - Ethics
 - Code of Ethics, 375
 - in engineering education, 374
 - morals and human behavior, 375
 - professional, 374–375
 - as subject in an undergraduate program, 376
 - as teaching methodology, 376–377
 - VSD, 374–376
 - European General Data Protection Regulation (GDPR), 85
 - European infrastructures, 182
 - Evapotranspiration (ET_o), 35
 - Evidence-based medicine (EBM), 179
 - Evolutionary algorithm (EA), 200
 - Evolution of viruses, 50
 - Explicit rating, 490
 - Exponential strategy, 266
 - Expression detection, 464
 - Extended Cohn-Kanade dataset, 465
 - Extended IFML (E-IFML), 274
 - Extended Yale B face, 465
 - Extensible Markup Language (XML), 480, 520–524
 - Extreme Learning Machine (ELM), 34, 35
- F**
- Facial expression recognition
 - BU-3DFE, 465
 - CNNs, 464, 467
 - compound emotion (CE), 465
 - computer vision techniques, 464
 - conventional approaches, 464
 - cutting edge technology, 464
 - deep-learning based, 464, 468
 - detecting process, 467
 - DISFA, 465
 - expression detection, 464
 - extended Cohn-Kanade dataset, 465
 - extended Yale B face, 465
 - facial landmark, 466
 - feature extraction, 466
 - fingerprint recognition, 464
 - flow chart, developed model, 467
 - healthcare applications, 468
 - healthy and sick persons, 464, 468
 - iris recognition, 464
 - JAFFE database, 464, 465, 467
 - KDEF, 465
 - MUCT landmarked face database, 465, 467
 - MUG facial expression database, 465–467
 - pandemic, 464
 - training accuracy, 467
 - training dataset, 465–466
 - training loss, 467
 - validity accuracy, 467
 - validity loss, 467
 - Faculty development, 506, 508
 - Feature descriptors, 471, 473
 - Feature extraction, 459, 466, 470–472
 - Feature filtering, 472
 - Feature importance, 505
 - Financial software, 86
 - See also* Software Requirements Analysis Method for Improvement of Privacy and Security (SRAM-PS)
 - First person point-of-view, 19
 - Fisher vector (FV), 471–474
 - 5G, 430, 433
 - Flight maneuvering, 19–24
 - Flight simulators, 19–24
 - Flu-like facial symptoms, 464
 - Foot vertices, 249
 - Forest Regression, 34
 - Formal method, 313
 - Formal verification, 256
 - 4-D architecture, 94
 - “Fourth Industrial Revolution,” 40
 - Frauds
 - in multiple choice exams, 207
 - in public tenders, 207
 - Fuzzy logic, 13, 15, 200
 - FV SIFT descriptor-based algorithms, 475
- G**
- Galois field arithmetic, 132
 - Game2Learn, 292
 - Game programming, 292
 - Game theory, 13, 294, 486
 - Gated Recurrent Unit (GRU), 222, 225
 - Gauss-Jacques method, 123–125, 128, 130
 - Gauss-Jordan with explicit modularization, 129, 130
 - Gender, 498–501
 - Gene expression, 403, 404
 - Generation of security, 82, 138, 142
 - Generative adversarial networks (GANs), 446–450
 - Genomics, 181–183
 - Geographical information system (GIS), 247, 501
 - and BIM, 351
 - computational system, 347
 - geographic space, 347
 - internet, 347
 - in web environment (GIS-Web) (*see* Web GIS)
 - Geometric mean, 516
 - Georeferenced data
 - mixed data (2D and 3D), 350–351
 - research data, 348
 - research questions, 348, 352
 - systematic review, 349, 351–352
 - 2D data, 349–350
 - visualization, 348

Geospatial data, 351, 352, 480, 500
 Gradient Boost Regressor, 34, 37
 Graph generation algorithm, 255–262
 Graphical user interface (GUI), 5, 294
 Graphics, 294
 Graphics processing units (GPU), 332
 ANN, 323
 ASNNs, 324
 comparison and speedup, 327, 329
 conventional feed forward, 323, 324
 CPU, 324
 CUDA, 325–326
 deep neural networks, 324
 layer, 324
 nodes, 324
 OPENCL, 325
 parallel GPU activation, 326–327
 sensor nodes, 323
 sequential activation, 326
 sequential results, 327, 328
 single GPU result, 327, 328
 sparse and arbitrary structured neural networks, 324
 Graph processing, 325
 Grazing visibility, 251–252
 Green computing, 267–270
 Green energy, 267–270
 Grey literature, 180
 Group Method of Data Handling (GMDH) algorithm, 30
 GUISurfer, 284

H

Hand-crafted features, 470, 473
 Hand-held devices, 265
 Hardware implementation
 algorithms, 307
 execution and data extraction, 308
 execution time, 307
 FPGAs, 309
 GPUs, 309
 implemented algorithms, 309–310
 performance and energy efficiency, 310–311
 research definition
 data extraction, 308
 execution, 308, 309
 questions, 308
 systematic mapping, 307–308
 Harmonic mean, 516
 Hash, 107–110, 112, 114
 HDCaml
 hardware circuit design, 316
 high-level design, 316–318
 Healthcare, 398
 applications, 468
 Healthcare Organizations, 160
 products, 148, 150–152
 Health infrastructure, 181, 182
 Health status monitoring and mental health applications, 482
 Heat map, 207, 349, 350
 Herd Immunity Threshold (HIT), 392
 Hermeneutics, 362
 Heuristics, 231, 362, 515
 Hibernate state, 265
 Hidden Markov Models (HMMs), 222
 Hierarchical Petri net Simulator (HIPS) tool
 analysis functions, 255, 257

 state space generator, 257
 Higher education, 503
 Higher Education Institutions (HEIs), 373
 High-tech AAC-devices, 170–171
 Hill cipher, 121–124, 128
 Home care, 164
 Home electronics, 430
 Home environment, 164
 Human-in-the-loop, 20
 Human systems thinking, 356
 Human trafficking
 bitcoin, 498
 blockchain, 498
 data, 498
 drugs and arms, 497
 machine learning, 498–501
 victims, 497
 Human values, 373, 374
 Hybrid model, 186
 Hybrid security risk assessment model
 BBN topology, 80–81
 CAPEC tools, 77
 CWSS, 139
 Dondo's approach, 78
 experience-based data, 77
 implementation, NVD files, 79
 Risk Estimation Model, 78–79
 risk management, 77
 sensitivity analysis, 81
 Hype Cycle, 428, 429
 Hyperledger, 417
 Hyperledger fabric, 417, 418
 HyperNEAT, 324

I

ICGC Argo project, 181
 IEEE (Institute of Electrical and Electronic Engineering),
 374
 guidelines, 215
 Xplore, 12
 I-I algorithm, 248–250
 Image analytics, 485
 Image classification, 470, 475
 Image instance retrieval
 CNNs, 474, 475
 fisher vectors, 474
 results, 474
 Image processing, 445, 449
 Image retrieval, 472, 474
 Image segmentation, 463
 Image-to-image translation, 445, 446
 Immersive virtual reality (IVR), 361, 362, 364
 ImpactCS project, 376, 378
 Implicit rating, 490
 Imputation methods
 diffusion smooth-based methods, 404
 single-cell datasets, 406
 statistical-based methods, 404
 Incremental Virtual Learning (IVL), 206
 Industrial IoT (IIoT), 148, 430
 Industry 4.0 (I4.0), 28, 30, 31
 Informatics, 377, 379
 Information and Communication Technologies (ICTs), 209
 advent, 212
 use, 211

- Information quality (IQ)
 - aggregation, 510
 - analytic hierarchy process, 510
 - assessment, 510
 - combining dimensions, 515–516
 - description, 509
 - education industry, 509–510
 - frameworks, 509, 511–515
 - identifying, 510–511
 - measurements, 510, 511
 - multiple ratings, 509
 - ratings, 510
 - statistical profiles, 510
 - theories, 509
 - Information security, 85–89, 127
 - Information technology (IT), 28, 265
 - multi-stakeholder PE model
 - Healthcare Organizations, 160
 - medical providers, 160
 - technology-focused PE, 159–160
 - in PE
 - care process, 155
 - electronic health interventions, 155
 - multi-actor model, 156
 - scoping review, 156
 - INRIAHolidays* and *UKBench* datasets, 474
 - Integrated development environment (IDE), 424
 - Integrate I4.0 technologies, 30
 - Intelligent decision support systems (IDSS), 3
 - Intelligent health care, 383
 - Intelligent modeling, 355
 - simulators and neural network frameworks, 358
 - solving pedagogical problems, 356
 - stages, 356
 - Intelligent systems, 355, 356, 358
 - See also* Artificial neural networks (ANNs)
 - Intel Turbo Boost Technology, 299
 - Internet, 69, 435, 513
 - Internet of Things (IoT), 385, 398
 - access control models, 416
 - aggregation, 431
 - agricultural applications, 34
 - applications, 433 (*see also* IoT applications)
 - big data, 428
 - cloud, 432
 - comprehensive study, 150
 - data collection, 431
 - edge devices, 427
 - emerging technologies, 413
 - gathering data, 432
 - inclusion and exclusion criteria, 414
 - integration, 431
 - interchange data, 428
 - organizations, 416
 - Python's support, 173
 - quality evaluation, 414
 - research process, 414
 - research questions, 413, 416
 - RFID, 428
 - smart city, 433
 - smart contracts (*see* Smart contracts)
 - smart environment, 430
 - Internships, 367–371
 - Interpersonal Reactivity Index (IRI), 363
 - Intrusion detection systems (IDS), 117
 - IoT applications
 - deployment strategy, 147
 - distribution process, 150–151
 - implementation, 147, 150
 - in SCM
 - impacts, 148
 - innovation, 148
 - maturity levels, 148–150
 - traditional management system, 148
 - supply chain, 147–149
 - iPad
 - and AAC applications, 171, 176
 - from Apple, 172
 - touch experience, 172
 - IQ, *see* Information quality (IQ)
 - IQ framework
 - accessible, 513
 - attractiveness, 511
 - credible, 513–515
 - definition, 511
 - desirable, 513
 - findable, 513
 - interaction, 511
 - trendiness, 511
 - usable, 512–513
 - useful, 512
 - UWebIQ, 511
 - valuable, 515
 - WebIQ, 513–515
 - Irrigation, 33–38
 - Item-item collaborative filtering, 342
 - Iterative heuristic process, 362
- J**
- Jaccard index (JI), 406, 407
 - Japanese Female Facial Expressions (JAFFE), 464, 465
 - Java Cryptography Extension (JCE), 424
 - Judge contract (JC), 416
 - Just in Time Compilers (JIT), 229
- K**
- The Karolinska Directed Emotional Face (KDEF), 465
 - Key matrix, 122–125, 128–130
 - K-Nearest Neighbor (KNN), 460, 461, 471
 - kNN-smoothing, 404
 - regression, 504
 - smart agriculture, 34
 - Knowledge base (KB), 3
 - Knowledge Sources of Security and Privacy Requirements (KSSPR), 87
- L**
- Labeled transition system (LTS), 257
 - Labor trafficking, 498
 - Latency, 438, 439
 - Law enforcement, 497, 498, 500, 501
 - Learning management system, 515
 - Linear regression, 79, 80, 82, 404, 445–450, 504
 - Linear support vector regression, 504
 - Linux kernel
 - Adaptive-ticks timer, 298, 299
 - APICs, 297
 - application optimization
 - highly CPU intensive threads, 303

- Linux kernel (*cont.*)
 - Inter thread communication, 303
 - lightweight threads, 303
 - management tasks, 303
 - multit-headed application, 303, 304
 - synchronization, 303
 - CPU cycles, 298
 - Dyntick-idle, 304, 305
 - experimental design, 299–300
 - GCC version, 301–302
 - high-performance computing, 298
 - Livermore loops, 300, 301
 - methodology, 299–300
 - multiple tasks support, 304
 - MySQL, 302
 - operating systems, 297
 - power consumption, 297–299
 - Power6 system, 298
 - RAMspeed, 300–301
 - settings, 299
 - SysBench, 302
 - system configurations, 299
 - timer tick frequency, 298
 - types, 297
- Logic library, 314–318
 - Login data, 107
 - Log4j java package, 422
 - Long Short Term Memory (LSTM), 222–223, 225
- LOTOS
 - abstract data types, 314
 - circuit module, 321
 - DILL, 313, 314
 - formal verification, 313
 - high-level description languages, 314
 - Internal Connecting Lines, 318
 - logic elements, 314
 - logic gates, 314–316
 - mathematical modeling languages, 313
 - multiplication, 320
 - n-bit parallel multiplier, 318, 319
 - nested structure, 321
 - parallel multiplier, 318–320
 - primary logic gate, 313
 - process synchronization, 318
 - repetitive circuit configuration, 316
 - RTL circuit description, 318
 - system design, 313
- Lower-division curriculum
 - academic institutions, 291–292
 - attributes
 - American children, 293
 - components, 293
 - drive learning, 294
 - effects of programming, 293
 - game software, 294
 - industry/research projects, 293
 - motivate students, 293
 - students' program, 293
 - Big Data, 294
 - educators, 293
 - game programming, 292
 - global education emergency, 291
 - instructional and educational research, 291
 - mathematics, 291
 - parallel algorithms, 293
 - robotics, 292
 - simple 2D/3D graphics libraries, 292–293
 - systematic review, 292
- Low tech AAC devices, 170, 171
- ## M
- Machine learning (ML), 205–206, 385, 498–501
 - algorithm classifications
 - reinforced learning, 40
 - supervised learning, 40
 - unsupervised learning, 40
 - application, 503–504
 - on cloud services, 39
 - CNN, 469
 - computer science, 463
 - computer systems, 11
 - computer vision, 469
 - correlation calculation, 505
 - CycleGAN, 449, 450
 - data extraction, 504
 - data preprocessing, 504
 - dataset, 446, 505–506
 - Decision Trees, 16
 - deep learning, 469
 - definition, 40
 - economic growth rates, 40
 - experimental setup, 506
 - feature importance calculation, 505
 - feature relevance, 506, 507
 - GANs, 446, 447, 449, 450
 - health care departments, 463
 - identification, 505
 - image processing, 445, 449
 - image-to-image translation, 445, 446
 - market analysis, 39–40
 - model evaluation and selection, 504–505
 - motivating research, 504
 - multi-output regression, 504
 - OpenCV, 446
 - overall satisfaction, 506, 508
 - pattern extraction, 457
 - pattern recognition, 451
 - Pix2Pix, 446, 448, 449
 - polaroid dataset, 447, 449
 - popularity, 39–40
 - predicted data set, 449
 - predictive accuracy, 506
 - preliminaries, 504
 - problem statements, 504
 - regression model, 446–450
 - RGB values, 447
 - smart agriculture, 34
 - SMR, 12
 - solutions, 39
 - and traffic control system, 11, 16
 - uncertainty, 451
 - unified vision proposal, 41–45
 - Machine learning tool, 4
 - MAGIC, 405, 406
 - Makeup, 451, 455
 - Malware
 - analysis tools, 56
 - BWSS, 422
 - CARO nomenclature, 51
 - classification, 50
 - on computer worms, 51, 52, 56

- evolution, 50
 - Internet, 54
 - strings, 52
 - Management information base (MIB), 432
 - Manual backup, 423
 - Map based, 348
 - Mapping study (MS), 12, 163–167
 - MapServer (map server), 350
 - Marketing, 433
 - Markov Decision Processes (MDP), 16
 - Mathematical modeling, 388
 - MATLAB simulator, 35
 - Maximum-Entropy Markov model (MEMM), 224
 - MDD, *see* Model Driven Development (MDD)
 - MD5 hashes, 108, 110
 - Mean average precision (mAP), 473–474
 - Mean squared error, 506
 - MEASUR (Methods for Eliciting, Analyzing and Specifying User Requirements), 86
 - Media Access Control (MAC) protocol, 92–95
 - Medical devices, 171
 - Medical imaging, 247
 - Medical services, 69, 74, 75
 - Mega floating-point operations per second (MFLOPS), 300
 - Memory efficiency, 255–262
 - Mentorships, 367–371
 - Merkle tree
 - Angela, 109
 - in C++, 107
 - data structure, 107, 108, 112
 - online sealed-bid auction, 108
 - parallel implementation, 111, 112
 - program functionality, 110
 - program structure, 110
 - type of application, 107
 - Message Passing Interface (MPI), 108, 109, 114
 - Metadata, 498, 501
 - Meta-heuristics, 13
 - Meta-learning supervisor neural network, 455
 - Methodology
 - for research, 374
 - VSD, 374
 - Microsoft Excel software, 28
 - Mining attack, 422
 - MMLib library, 490
 - Mobile applications, 215, 401
 - Mobile application testing
 - analysis, 274
 - Android development kit, 274
 - Android Ripper, 274
 - categories, 274
 - CCov generation, 276–278
 - combinatorial-based sequence criteria, 273
 - devices, 273
 - efficiency, 274
 - efficient testing tools, 273
 - element sequences, 273
 - experimental setup, 276
 - generation algorithm, 274–276
 - generation-based test, 274
 - MobiGUITAR, 274
 - open-source applications, 274
 - random generation, 277
 - random walk, 276, 277
 - Scov generation, 276, 277
 - threats to validity, 278
 - Mobile technologies, 398
 - Mobile Usability Smell Evaluator (MUSE), 284, 286
 - Model Driven Development (MDD)
 - and agile development, 519
 - transformation mechanisms, 520
 - WebAC-MDD (*see* Web Agile and Collaborative Model-Driven Development (WebAC-MDD))
 - web application, 520
 - Modeling language, 519–524
 - Model Specific Registers (MSRs), 300
 - Model transformation, 520–524
 - Model View Controller (MVC), 351, 525
 - Modern-day slavery, 497
 - See also* Human trafficking
 - Modular inverse matrix, 123, 124, 128, 129
 - Monoalphabetic ciphers, 121
 - Monotone mountain, 251
 - Monotone polygon, 251
 - Morville’s User Experience Honeycomb, 511
 - Motivated Strategies for Learning Questionnaire (MSLQ), 292
 - Motivating students, 291, 292, 294
 - Motivating topics, 292–294
 - Motivational science, 293
 - MUCT landmarked face database, 465
 - MUG facial expression database, 465, 466
 - Multi-Criteria Decision Analysis (MCDA), 197
 - Multi-Criteria Decision-Making (MCDM), 197
 - Multi-hop network, 92
 - Multi-output regression, 504
 - Multiple encryption, 123–125
 - Multiple matching removal (MMR) algorithm, 473
 - Multivariables, 60, 61
 - Multivariate-polynomials, 133
 - Myers-Briggs theory, 457, 459
 - Myers-Briggs Typological Indicator (MBTI)
 - combining classifiers, 458, 461
 - feature extraction techniques, 459
 - judging functions, 458
 - perception functions, 458
 - personalities, 458
 - psychometric scheme, 458
 - MySQL, 302
- ## N
- NADTW (New Approach for Detecting TCP worms), 51
 - Naive Bayes (NB), 460, 461
 - algorithm, 458
 - Named Entity Recognition (NER), 225
 - Narrative review method, 60, 66
 - Narrow Band Internet of Things (NB-IoT), 436
 - National Institute of Standards 98 and Technology (NIST) 800-53, 60, 61, 78, 138
 - National Vulnerability Database (NVD), 77–82
 - Natural language processing, 222
 - Needleman-Wunsch algorithm, 310
 - NetBeans, 424
 - Network attack, 422
 - Network load (NL), 92, 93, 95–100, 102
 - Networks, 413
 - VLANs (*see* Virtual Local Area Networks (VLANs))
 - Network size (NS), 92, 95–100
 - Neural Evolution of Augmenting Topologies (NEAT), 324
 - Neural networks, 224, 404, 405, 467
 - See also* Artificial neural networks (ANNs)
 - Neuroscience, 355

New processor architectures, 131
 NEWS–FAST–COVID, 399
 NEWS2 protocol, 399, 401, 402
 Next-generation sequencing, 179
 NIST's Post-Quantum Cryptography (PQC) Standardization project., 135
 NIST Web Metrics, 511
 Nodes, cybersecurity analysis
 categorization of vulnerabilities, 73
 CVE data, 73, 74
 DICOM protocol, 70
 evaluation of data, 74–75
 experimental design
 categorization of services, 72
 discovery of nodes, 72
 discovery of services, 72
 discovery of vulnerabilities, 72
 vulnerability analysis, 72
 vulnerability validation process, 73
 PACS servers, 69
 servers/operational nodes, 73
 WADO communication methods, 73
 Non-Steiner vertices, 252
 Normalized mutual information (NMI), 406, 407
 Noyce summer workshops, 367–371
 NSF Robert Noyce Teacher Scholarship program, 371
 NVD Key Term, 79
 NVIDIA System Management Interface, 335

O

Obfuscation, 50, 52–54, 56
 OCaml, 314, 316, 317, 321
 Occlusion, 451, 455
 Offline algorithms, 266
 Offline strategies, 267, 268
 “Omics,” 179
 1-D architecture, 94
 One-time password (OTP), 425, 426
 Online algorithms, 265–268
 Online competitive analysis, 265
 Online instruction, 370, 371
 Online learning model, 16
 Online project management, 367, 369–371
 Online strategies, 267, 268
 On-the-fly removable states detection, 258
 Ontologies, 3–4
 breast cancer prediction system, 4–5
 CRO, 4
 OWL, 4, 5
 shared domain, 4
 symptom ontology, 5
 vulnerabilities and attacks on VLAN (*see* Virtual Local Area Networks (VLANs))
 OpenCL, 324
 Open data, 212
 OpenDill, 314–318
 OpenMP (C++ library), 108–109
 Open-source, 172, 173, 175, 176
 Open Source Intelligence (OSINT), 498
 Optocoupler, 437
 Organizational semiotics, 86, 87
 Original vertices, 249
 Oversampling techniques, 460

P

Packet Delivery Ratio (PDR), 93, 95–102
 Packet size (PS), 92, 93, 95–101
 PACS servers (Picture Archiving and Communication System), 69, 70, 72, 74
 Parallel computing, 185, 186, 293
 Paralleling generators, 255, 256, 258, 260–262
 Parallel programming, 110
 Parrot AR 2.0 drone, 19–21
 Particle simulation, 186
 Part-Of-Speech (POS), 461
 Password
 Merkle tree, 110
 parallel implementation, 110–111
 program functionality, 110
 program structure, 110
 Password-based authentication, 425
 Patient engagement (PE)
 approaches and interventions
 biotech companies, 158–159
 clinical providers, 157–158
 E-health solution developers, 157
 insurers (payors), 157
 engaging patients, 155
 involvement strategies, 156–157
 IT role (*see* Information technology (IT))
 multi-stakeholder PE model
 Healthcare Organizations, 160
 medical providers, 160
 technology-focused PE, 159–160
 passive and active engagement, 155–156
 patient activation, 156
 patient insight, 156, 158, 159
 Patient Information Based Algorithm (PIBA), 31
 Patient Measurement Analyzer Service, 401
 Pattern recognition, 247, 451
 Payors, 157, 159, 160
 Pedagogical empirical experience, 356
 Pedagogical problem, 356–358
 Pedagogical task/situation, 356
 Peer to peer (P2P), 416
 People with visual impairment (PVI), 163, 165, 166
 Performance, 298, 299
 Performance analysis, 491, 492, 495
 Performance comparison, Petri nets
 experimental conditions, 260
 single/multi-thread implementation, 261–262
 TCPcondis, 260–261
 Permissioned blockchain, 418, 419
 Personality, 458, 461
 Personality Cafe, 459
 Personalized precision medicine (PPM), 179, 180
 “Perspective” project assignment, 358
 Perspective taking (PT), 361–363
 See also Virtual reality perspective taking (PTVR)
 Pervasive home health care, 398
 Petri nets
 asynchronism, 255
 HiPS (*see* Hierarchical Petri net Simulator (HiPS) tool)
 implemented, 256
 incidence matrix, 256
 nondeterminism, 255
 performance comparison, 260–262
 properties, 255
 P/T-net, 256

- reachability, 256–257
 - removable states detection, 258–260
 - state equation, 256
 - state space generation, 255
 - state space generator, 257–258
 - Pharmaceutical companies, 158
 - Physical money, 421
 - Picture Exchange Communication System, 170
 - Piece-wise linear function, 265–271
 - Pix2Pix, 446, 448, 449
 - Place/transition net (P/T-net), 256
 - Plaintext, 121–125
 - Playfair cipher, 121, 122
 - Polaroid, 445, 447, 448, 450
 - Polyalphabetic ciphers, 121
 - Polygonal approximation, 247–249
 - Polygonal chains, 247
 - Portfolio balancing
 - management, 195
 - selection and prioritization, 195–196
 - Power down
 - analysis of changing
 - duration with fixed slope, 267–269
 - slope with fixed duration, 269–270
 - electrical energy, 265
 - exponential strategy, 266
 - hand-held devices, 265
 - hibernate state, 265
 - information technology, 265
 - online algorithms, 266
 - online and offline strategies, 267
 - online competitive analysis, 265
 - piece-wise linear function, 267
 - power states, 265–266
 - power usage, 265
 - prior work, 266
 - suspend state, 265
 - Power states, 265–266
 - Power usage, 265
 - PRECIMED project, 33
 - Predictive accuracy, 506
 - Predictive analytics, 485
 - Preliminary diagnosis, 177
 - Prescriptive analytics, 485
 - Privacy
 - regulations, 8
 - and security, 85, 86
 - SRAM-PS, 85–89
 - Problem Articulation Method (PAM), 86
 - Processors, 107–110, 112, 114
 - Product innovation, 430
 - Professional development, 367, 368
 - Professional ethics, 374–375
 - Programming in Python, 355, 358
 - Project selection, 195
 - PROMETHEE, 198
 - Propagation, 50, 51, 53–56
 - Protege (open-source software)
 - breast cancer ontology
 - basic query in DLQuery, 6, 8
 - classes and subclasses, 5, 6
 - data property hierarchy, 6, 7
 - determination of domain, 5
 - implementation process, 5
 - object property hierarchy, 6, 7
 - OntoGraf tool, 5, 6
 - OWLViz visualisation, 5, 7
 - SPARQL query language, 6
 - validation, 6
 - Protocol
 - DICOM (*see* DICOM (Digital Imaging and Communications in Medicine) protocol)
 - MAC protocol, 92–95
 - Public blockchain, 418, 419
 - Public participation
 - policies, 210
 - Pythagorean theorem, 511
 - Python, 173
 - Python libraries, 506
 - Python programming language, 173, 334, 355, 358
- Q**
- Q-learning
 - deep Q-learning, 13, 15
 - MDP, 16
 - model-free, 16
 - RL models, 13–16
 - Quality of service (QoS), 91, 92, 94–96, 100
 - Questionnaires, 5, 7, 216, 218, 219
- R**
- Radial basis function (RBF), 504
 - Rainbow (signature schemes), 131–135
 - RAMspeed, 300–301
 - Random Forest, 504
 - classifier, 500, 501
 - regressor model, 241
 - smart agriculture, 34, 37
 - Random oversampling, 461
 - Raspberry Pi, 165, 172–176, 417
 - Rayleigh component, 331
 - RBF, *see* Radial basis function (RBF)
 - Reachability, 256–257
 - Reachable graph, 258–260
 - Recommender Systems Evaluator (RSE)
 - applications, 339
 - classic algorithms, 340
 - collaborative filtering approach, 339
 - database, 340
 - experimental results
 - assay, 343, 344
 - environment, 342
 - factor influence, 343, 344
 - item-item collaborative filtering., 342
 - relevance specification, 342
 - response variable, 343, 344
 - RMSE, 343
 - setup, 342
 - test data, 342
 - throughput, 343
 - information and clients, 339
 - Lenskit*, 340
 - Netflix, 341
 - optimal parameter values, 340
 - performance evaluator
 - configuration file, 341
 - evaluation workflow, 341, 342
 - factor and levels, 341
 - load distribution, 342
 - reports, 342

- Recommender Systems Evaluator (RSE) (*cont.*)
 response variables, 341
 stochastic process, 341
 PostgreSQL, 340
RiVal, 340
 software, 340
 statistics, 340
TagRec, 340
 traditional approaches, 339
 workflow, 340
- Recurrent neural networks (RNNs), 16, 34, 222, 464, 486
- Refraction microtremor (ReMi) surveys, 331
- Register contract (RC), 416–417
- Register Transfer Level (RTL), 316
- Regression
K-nearest neighbors, 504
 linear, 504
 linear support vector, 504
 models, 241–242, 446–450
 multi-output, 504
 random forest, 505
- Regression task, 240–241
- Reinforcement Learning (RL), 40
 application, 16
 approaches, 15
 classification, learning articles, 15
 dominance, 13
 meta-heuristics, 13
 predominance, 13
- Relational Database Management Systems (RDBMS), 484
- Relaxed version, 252
- Reliability, 413, 429
- Remote learning, 368–371
See also E-NEST remote learning
- Remote monitoring technology, 401
- Remote MySQL database, 223
- Remote Patient Monitoring (MRP), 398, 399
- Removable states detection
 conditions, 258–259
 and detection parallelization, 260
 identifying, 259–260
 on-the-fly, 258
 reachable graph, 258–260
- Renewable energy, 265
- Requirements engineering (RE), 86, 88, 216, 219
- Resampling techniques, 457, 460
- Research and Development (R&D) programs, 199
- Retail powerhouse platform, 433
- Risk assessment
 hybrid security model (*see* Hybrid security risk assessment model)
- Risk estimation model, 78, 137
- Risk management, 137
 and sensitivity analysis, 141–143
- Robert Noyce Teacher Scholarship program, 368
- Robotics, 292
- Role-based access control (RBAC), 416
- Root-mean-square error (RMSE), 343
- Root node, 107, 108, 110
- RSA cryptosystem, 422–426
- “RUDO: A Home Ambient Intelligence System for Blind People,” 164
- Running-time, 107–114
- S**
- Sale process, 150–152
- Sans-Enterprise Wireless Checklist, 60, 61
- SARS-Cov-2
 cause, 28
 learning algorithms, 28
 outbreak, 27, 388
 social distancing policies, 390, 395
- SARS-CoV-1 epidemics, 27
- SAVER (scRNA-seq), 404–406
- SCADA (Supervisory Control and Data Acquisition) systems., 69, 117
- Scalable systems, 489, 490
- Scale-invariant feature transform (SIFT), 470, 472–475
- Scene Layer Package (SLP)., 352
- Scheduling timers, 297, 298, 304
- Scikit-learn, 500
- scImpute models, 404, 406
- Scopus, 12
- scRecover, 404
- Search Engine Optimization (SEO), 513
- Secret key, 121, 128, 133, 423
- Secure audit log (SAL), 424
- Secure communication among services, 401
- Security, 414–416
 network security services, 116
 of web applications, 118
- Security information, 127
See also Cryptography
- Security variables in wireless networks
 best practices, 61
 conceptual model, 66
 multivariates, 60, 61
 reference books, 61, 63–64
 research security variables, 61, 62
 standards, 60–61
 studies, 60
 users, 65
 variables result, 65
 Venn diagram, 65, 66
- SEEQ, *see* Students’ Evaluations of Educational Quality (SEEQ)
- Segmentation, 116, 326, 329, 463, 470
- Seismic refraction, 331
- Self-replicating program (SRP), 49
- Semantic medical profile, 5
- Sensitivity analysis, 81–83, 141–143
- Sensor-cloud, 432
- Sensor-cloud Sensing-as-a-Service (SSaaS), 432
- Sensor networks, 432
- Sensors, 398, 432
- SEO, *see* Search Engine Optimization (SEO)
- Sequence alignment, *see* Hardware implementation
- Sequence-based combinatorial coverage (SCov), 275
- Serial programming, 111
- Servers, 437
 Apache HTTP server, 75
 general data of nodes, 74
 operational nodes, 73
 PACS-DICOM servers, 70, 74
 PACS server, 69, 70, 72
 with vulnerabilities, 73
- Service-Based Architecture (SOA), 401
- SET, *see* Students’ evaluations of teaching (SET)
- Severe Acute Respiratory Syndrome (SARS), 27, 397
- Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), 397
- Shannon’s perfect secrecy, 123, 125
- SigFox, 401
- Signature schemes, 131–133
See also Rainbow (signature schemes)
- Simple 2D/3D graphics libraries, 292–293

- Simple polygon, 251–252
- Single-cell Imputation using Residual Network (scIRN)
 - DrImpute, 405
 - expression matrix, 404
 - imputation methods, 404, 405
 - imputing dropout data, 404–405
 - low-dimensional representation, 404
 - MAGIC, 405
 - missing data, dropout events in scRNA-seq, 409
 - modules, 404
 - neural network, 404
 - non-redundant representation, 404
 - SAVER, 405
 - single-cell datasets, 404–406
 - sub-population identification, 406, 409
 - transcriptome landscape
 - Klein dataset, 408
 - UMAP, 408, 409
 - visualization, 406–409
 - workflow, 404, 405
- Single-cell resolution, 403
- Single-cell RNA sequencing (scRNA-seq)
 - challenges, 403
 - dropout challenge, 403
 - dropout events, 403
 - gene expression, 403, 404
 - imputation methods, 403–404
 - scIRN (*see* Single-cell Imputation using Residual Network (scIRN))
 - zero expression values, 403
- Singledimension, 516
- Single-hop network, 92
- Single/multi-thread implementation, 261–262
- Single-thread search algorithm, 257–258
- SIR (Susceptible, Infected, Recovered) model, 387–399
- SIRV model, 387, 392, 395
- Slant stack transform
 - approach/implementation, 333
 - CUDA, 332–333
 - Fourier transforms, 335
 - GPU memory, 335
 - graphics processing units (GPU), 332
 - hardware, 334
 - machine learning model, 331, 335
 - Nyquist frequency, 332
 - 1D shear-velocity vs. depth model, 331
 - parallel process, 335
 - p-f plots, 331, 332
 - p- τ plots, 332, 335
 - processor utilization, 335
 - PyCUDA, 334
 - ReMi, 331
 - software, 334
- Smart agriculture, 33–34
 - See also* CropWaterNeed approach
- Smart city, 209, 433
 - and corporate interests, 213
- Smart contracts
 - ACC, 416
 - architecture, 438
 - blockchain, 416, 417
 - blockchain network, 437
 - definition, 437
 - distributed applications, 416
 - Ethereum blockchain, 437
 - hyperledger, 417
 - hyperledger fabric, 417, 418
 - IoT, 417
 - JavaScript language, 437
 - judge contract, 416
 - limitations, 418
 - network, 416
 - public blockchain scalability, 418
 - register contract, 416–417
 - writing data, 438
- Smart environments, 164, 413, 414, 430
- Smart grid, 265
- Smart home, 164, 430
- Smart irrigation, 33
- Smart lock, 416
- Smart meters, 437
- Smart power meter, 436
- Smart regression, 34, 36, 37
- Smart spaces, 163
- Smart supply chains, 147
- Smith-Waterman algorithm, 310, 311
- SMSLib java package, 422
- Snowballing, 12, 205
- Social media
 - active users, 484
 - active user statistics, 483
 - applications, 433
 - BDA, 479, 484–485
 - big data, 480
 - big social data, 481, 483–484
 - businesses, 427, 431, 432, 482
 - CMO, 479
 - criteria, 481
 - data science research, 482
 - Enterprise 2.0, 482
 - health status monitoring and mental health applications, 482
 - information sharing instant and speedy, 479
 - integration, 432
 - multi communication, 481
 - public and emergency responders, 479
 - 60 seconds, 482
 - Society 2.0, 481
 - statistics, 483
 - types, 480, 481
 - websites/applications, 481
- Social media analytics, 484
- Social media-based big-data-driven decision systems, 479
- Social media data analysis, 480
- Social networking websites, 480
- Social networks, 427, 433, 457, 459
- Social Set Visualizer (SoSeVi), 485
- Society 2.0, 481
- Software, 376
- Software as a service (SaaS), 432
- Software development project, 519
- Software engineering, 215
- Software Requirements Analysis Method for Improvement of Privacy and Security (SRAM-PS), 85–89
- Solutions based
 - on data mining, 206
 - on statistics, 206
- Source code, 117, 234, 237, 281, 283, 450, 519–524
- Source-code generation, 520–524
- Spark the Definitive Guide*, 490
- SPARQL query language, 6, 8
- Species, 51, 54, 56
- Speech assistant, 169, 171, 176

- Speech Generating Devices (SGDs), 169
- Spoofing, 451
- Spread monitoring, UAV, 383–386
- SQLite, 223
- SSCHs, *see* Student-semester credit hours (SSCHs)
- Stacked Bayesian Self-learning Network, 404
- Standards
 - CIS controls, 61
 - ISO/IEC 27001, 60, 61
 - NIST 800-53, 60, 61, 78, 138
 - security variables, 61
- State space generation, 255–256
 - by parallel threads, 258
 - removable states detection, 258–260
 - by single thread, 257–258
- Statistical-based methods
 - bayNorm, 404
 - RIA, 404
 - SAVER, 404
 - scImpute models, 404
 - scRecover, 404
- Statistical Package for the Social Sciences (SPSS), 79
- Steiner vertices, 250–252
- Strategic alignment, 195, 196
- Streaming Multiprocessors (SM), 333
- Student-semester credit hours (SSCHs), 512
- Students' Evaluations of Educational Quality (SEEQ), 504, 505, 508
- Students' evaluations of teaching (SET), 504
- Substitution technique, 121
- Summer workshops, Noyce, 367–371
- "Sunflower Model of Big Data," 480, 481
- Supervised learning, 40
- Supply chain
 - challenges, 147
 - in health care products, 148
 - IoT and application, 147, 148
 - SCM, 147
 - traditional management systems, 148
- Supply chain management (SCM), 147
 - IoT application
 - impacts, 148
 - innovation, 148
 - maturity levels, 148–150
 - real-time SCM, 148
 - traditional management system, 148
- Support Vector Classification (SVC), 240
- Support Vector Machine (SVM), 240, 460, 461, 471
- Support Vector Regression (SVR), 34, 241
- Suspend state, 265
- Symmetric cryptography, 127, 128
 - See also* Cryptography
- Symmetric encryption, 121–123, 424
- Symptoms monitoring, 398–399
- Synthetic Minority Over-sampling Technique (SMOTE), 460, 461
- SysBench, 302
- System
 - EDOWA system, 50
 - Internet, 54
 - systematic review, 51
 - WDS, 50
- Systematic mapping, 201, 204
 - articles and data extraction, 196–197
 - definition, 196
 - exclusion criteria, 196
 - process, 204–205
 - research question definition, 196
 - screening process, 197
 - search strategy and search strings, 196
 - selected articles, 197
 - selection criteria, 196
 - techniques and tools, 197
- Systematic mapping review (SMR)
 - inclusion and exclusion criteria for publications, 12
 - research question, 12
 - search engines and search expressions, 12
 - trends and gaps, 16
- T**
- Taxonomies, 51–54, 56, 116–118
 - vulnerabilities and attacks on VLANs, 117, 118
- TCP connection procedure model (TCPcondis), 260–261
- Technologies, 39, 398
 - AAL, 163, 164, 166
 - convergence, 141
 - development, 378
 - exponential, 373
 - in human scenarios, 373
 - human trafficking (*see* Human trafficking)
 - law enforcement agencies, 498
 - RFID, 149, 165, 428, 430
 - social impact, 373
 - VSD (*see* Value Sensitive Design (VSD))
- Temporal variations, 464
- Term Frequency (TF), 459
- Term Frequency Inverse Document Frequency (TFIDF), 459
- Terrain illumination, 251, 252
- Terrain modeling, 251
- Terrestrial Wireless Sensor Networks (TWSNs), 92–94
- Test suite generation, *see* Mobile application testing
- Text analytics, 485, 501
- Text to speech processing, 169, 172–176
- Threading building blocks (TBB), 258
- Threads, 108, 109, 112, 114
- Threats
 - APT, 116
 - cyber, 69, 70, 73, 75
 - web applications, 118
- 3D architecture, 94
- 3D GIS, 349–351
- 3D traffic geographic information subsystem, 351
- Time-dependent SIR, 387
- Total Collision (TC), 97–102
- Traditional computer vision (CV)
 - CNN, 470
 - deep learning, 469–471
 - vs.* DL
 - CNN, 470–471
 - feature descriptors, 471, 473
 - fisher vector, 472–473
 - hand-crafted features, 470, 473
 - SIFT, 472
 - extracted hand-crafted features, 470
 - feature extraction stage, 470, 475
 - filtering criteria, 473
 - image dataset, 470
 - pre-processing stage, 470
 - selection process, 475
 - SIFT, 469, 473, 475
- Traffic control, 11, 12, 16
- Traffic light coordination
 - classification models, 239, 241

- classification task, 240
 - data set, 239–240
 - regression models, 239, 241–242
 - regression task, 240–241
 - Training predictive models, 503–505
 - Transaction latency, 439
 - Transaction response time, 439
 - Transportation research, 247
 - Trip status at a signal, 239
 - TrustZone technology, 422
 - 2D architecture, 94
 - 2D GIS, 349–350
 - Two factor authentication (2FA), 425, 426
- U**
- UAV-based system
 - for COVID-19 operations, 384
 - evaluation, 385
 - UDP packet, 20–22, 24
 - UKBench* dataset, 474
 - Unaided AAC-devices, 170
 - Underwater Acoustic Modems (UAMs), 95–96
 - Underwater acoustic sensor networks (UASNs)
 - challenges, UASN communication, 93–94
 - communication, 92
 - DRs, 96
 - environmental factors, 95
 - MAC protocol, 92–95
 - multi-hop networks, 92
 - network architecture, 94
 - network topology, 92, 93
 - performance evaluation, 96–99
 - sensors, 91
 - vs. TWSNs, 92
 - UAMs, 95
 - underwater applications, 91
 - Underwater sensors, 91, 95
 - Unified Modeling Language (UML) profile, 520, 524
 - Unified Parallel C (UPC++), 109
 - University of California-Berkeley, 295
 - University Website Information Quality (UWebIQ), 511
 - Unmanned aerial vehicles (UAVs), 20, 383
 - spread monitoring, 383–386
 - Unsupervised learning, 40
 - Usability smells
 - activities, 281
 - assessment, 281
 - bad smells, 283
 - bad usability smells, 281, 284–285
 - code smells, 281, 283, 284
 - comparison strategy, 286
 - efficient algorithms, 287
 - GUISurfer tool, 286
 - interactive applications, 287
 - interfaces, 281
 - level of usability, 281
 - planning
 - data extraction strategy, 283, 284
 - primary studies, 282
 - quality assessment, 282, 283
 - research questions, 282
 - search strategy, 282
 - software systems, 281
 - task-based, 284
 - techniques, 285
 - tools, 285
 - user satisfaction, 281
 - web applications, 281
 - Usability Smells Finder (USF), 284–286
 - User data, 489
 - User-data-driven recommendations, 490
 - User preferences, 490
 - UseSkill Extension (USE), 284
 - UWebIQ, *see* University Website Information Quality (UWebIQ)
- V**
- Vaccine distribution strategy, 387, 393–395
 - Validation dataset, 504
 - Value Sensitive Design (VSD), 374, 379
 - conceptual research, 375
 - Display Cards, 375
 - empirical investigations, 375
 - forecast cards, 375
 - on stakeholder analysis, 375
 - technical investigations, 375
 - Variables, 61
 - See also* Security variables in wireless networks
 - V-aware approximation problem (VAP), 249–250
 - Vectorized GFNI, 132
 - Vector tile strategy, 350–352
 - Vehicles
 - AI techniques, 11
 - Brazilian vehicle fleet, 11, 12
 - V2I, 16
 - Vehicle to Infrastructure networks (V2I), 16
 - Vehicular traffic systems, 11
 - Video analytics, 485
 - Virtual Hospital
 - biosensors, 401
 - clinical MRP, 400
 - Cloud Services model, 401
 - COVID-19, 398–399
 - data storage, 401
 - environments, 401
 - hospital environment, 401
 - IoT, 401
 - monitoring center, 399
 - monitoring situations, 401
 - MRP, 399
 - NEWS2 protocol, 399
 - pandemic, 397, 402
 - remote monitoring system, 399
 - remote patient monitoring systems, 398
 - virtual machines, 401
 - web application, 401
 - Virtual Local Area Networks (VLANs), 115–118
 - APT, 116
 - attacks, 115–118
 - network environments, 115
 - survey, 118
 - VoIP service, 116
 - vulnerability, 116–118
 - Virtual machines
 - API, 401
 - application, 401
 - brokers, 401
 - database, 401
 - DBMS, 401
 - mobile application, 401
 - patient Measurement Analyzer Service, 401

- Virtual machines (*cont.*)
 - secure communication among services, 401
 - web application, 401
 - Virtual reality (VR), 351
 - communication medium, 361
 - in education, 361, 363
 - and empathy, 361–362
 - environment, 362
 - IRI, 363
 - IVR, 361, 362, 364
 - literature review, 361–363
 - perspective taking (PT) (*see* Virtual reality perspective taking (PTVR))
 - simulations, 363
 - “Zonas VR,” 361
 - Virtual reality perspective taking (PTVR), 361, 363, 364
 - Visibility, 249–252
 - Visibility graph (VG), 251–252
 - Visibility polygon (VP), 249
 - Visibility preserving, 249
 - VisiTot, 252
 - Visual impairment, 164
 - See also* People with visual impairment (PVI)
 - Visualization, 406–409
 - See also* Georeferenced data
 - Visually observable diseases, 463
 - Voting combination approach, 458
 - “3 Vs” (Volume, Velocity and Variety), 480
 - Vulnerabilities, 59, 65, 70–72
 - CVSS, 78, 79
 - VLANs, 116–118
 - See also* Nodes, cybersecurity analysis
- W**
- Wallet damage, 422
 - W3C’s, 513
 - Web Agile and Collaborative Model-Driven Development (WebAC-MDD), 520–524
 - Web Agile Modeling Language (Web-AML), 520, 524
 - Web analytics, 485
 - Web application, 401
 - CHI page, 523
 - design, 520
 - development, 519
 - domain classes structure, 521, 522
 - MDD, 520
 - source-code, 520, 522–524
 - tests structure, 523
 - XML files, 520
 - WebAssembly
 - applications, 230
 - ASM.JS code, 229
 - backend machines, 231
 - backend servers, 237
 - client machines, 231
 - client-side, 230
 - CPU-intensive tasks, 229
 - device free memory, 231–232
 - device information, 231
 - evaluation
 - applications, 233
 - client-based, 232
 - decider algorithm, 232
 - decider time, 232, 233
 - JavaScript, 232
 - LGE phone vs. server, 237
 - load and insatiate, 233, 234
 - network time vs. total time, 232, 233
 - OpenCV application, 235
 - Ostrich benchmark performance, 232, 233
 - page load time, 234, 235
 - polyfill script, 232
 - server-based, 232
 - WASM/JS execution time, 234, 235
 - Zip time plus network time, 234, 236
 - ZipT time, 234, 236
 - Zstd time, 234, 235
 - Zstd time plus network time, 234, 236
 - experimental setup, 231
 - experiment design, 231
 - internet, 230
 - Java Applet, 229
 - JavaScript, 229, 237
 - ML model, 237
 - mobile browsers, 230
 - network, 232
 - system-like software, 229
 - telecommunication infrastructure, 230
 - web applications, 230
- Web content accessibility guidelines (WCAG), 513
- Web GIS, 347
 - 1D GIS, 349, 350
 - 2D GIS, 349–350
 - 3D GIS, 349–351
 - methodology, 348
 - research data, 348
 - research questions, 348
 - technologies, 347
- WebIQ
 - avoid even small errors, 515
 - avoid promotional ads, 515
 - contacting the author, 514
 - display the content expertise, 514
 - honesty and trustworthiness, 514
 - information accuracy, 513–514
 - organization of website, 514
 - professional, 514
 - useful and easily, 514
 - UWebIQ, 511
- WebML, *see* Web Modeling Language (WebML)
- Web Modeling Language (WebML), 520, 524
- Web Ontology Language (OWL), 4, 5
- Website Analysis and Measurement Inventory (WAMMI), 511
- Websites
 - accessibility IQ, 513
 - content expertise, 514
 - current, 514–515
 - and digital services, 511
 - display the content expertise, 514
 - Federal Agencies, 510
 - findable, 513
 - health information, 510
 - honesty and trustworthiness, 514
 - Namik Kemal University, 511
 - organization, 514
 - priority, 509, 510
 - requirements, 511
 - university, 509–517
 - usable, 512–513
 - useful and easily, 514
 - WAMMI, 511

- Web traffic, 489
 - Whole-genome sequencing (WGS), 180
 - Wi-Fi wireless networks, 59
 - analysis and data interpretation, 61, 65
 - best practices, 61
 - conceptual model, security variables, 66
 - phases of research, 60
 - Sans-Enterprise Wireless Checklist, 60, 61
 - security, 60 (*see also* Security variables in wireless networks)
 - selection of variables, 60
 - standards, 60 (*see also* Standards)
 - Wireless checklist
 - Sans-Enterprise Wireless Checklist, 60, 61
 - Wireless Sensor Networks (WSNs), 35, 432
 - W-learning, 13, 15
 - World Wide Web (WWW), 347
 - See also* Information quality (IQ)
 - Worm detection systems (WDS), 50
- X**
- XGBoost Regressor
 - precision agriculture, 34, 37
 - XGBRegressor model, 37
- Z**
- Zero-inflated negative binomial model (ZINB), 404
 - “Zonas VR,” 361