Studies in Big Data 89

Sujata Dash
Subhendu Kumar Pani
Ajith Abraham
Yulan Liang *Editors*

Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing

Springer

Studies in Big Data

Volume 89

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series "Studies in Big Data" (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are reviewed in a single blind peer review process.

Indexed by SCOPUS, EI Compendex, SCIMAGO and zbMATH.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/11970

Sujata Dash · Subhendu Kumar Pani ·
Ajith Abraham · Yulan Liang
Editors

Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing

Springer

Editors
Sujata Dash
Department of Computer Science
and Application
North Orissa University
Bhubaneswar, India

Subhendu Kumar Pani
School of Computer Science
and Engineering
Orissa Engineering College
Bhubaneswar, Odisha, India

Ajith Abraham
Faculty of Electrical Engineering
and Computer science
Machine Intelligence Research Labs (MIR Labs)
Auburn, WA, USA

Yulan Liang
Department of Family and Community Health
University of Maryland
Baltimore, MD, USA

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-3-030-75656-7 ISBN 978-3-030-75657-4 (eBook)
https://doi.org/10.1007/978-3-030-75657-4

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

Preface

Overview

The new applications of soft computing can be regarded as an emerging field in computer science, automatic control engineering, medicine, biology application, natural environmental engineering, and pattern recognition. Now the exemplar model for soft computing is the human brain. The use of various techniques of soft computing is nowadays successfully implemented in many domestic, commercial, and industrial applications due to the low-cost and very high-performance digital processors and also the declining price of the memory chips. This is the main reason behind the wider expansion of soft computing techniques and their application areas.

These computing methods also play a significant role in the design and optimization of diverse engineering disciplines. With the influence and the development of the Internet of Things (IoT) concept, the need for using soft computing techniques has become more significant than ever. In general, soft computing methods are similar to biological processes very closely than traditional techniques, which are mostly based on formal logical systems, such as sentential logic and predicate logic, or rely heavily on computer-aided numerical analysis. Soft computing techniques are anticipated to complement each other. The aim of these techniques is to accept imprecision, uncertainties, and approximations to get a rapid solution. However, recent advancements in the representation of soft computing algorithms (Fuzzy Logic, Evolutionary Computation, Machine Learning, and Probabilistic Reasoning) generate a more intelligent and robust system providing a human interpretable, low-cost, approximate solution.

Soft computing-based algorithms have demonstrated great performance in a variety of areas including Multimedia retrieval, Fault tolerance, System modeling, Network architecture, Web semantics, big data analytics, time series, biomedical and health informatics, etc. Soft computing approaches such as Genetic Programming (GP), Support Vector Machine–Firefly Algorithm (SVM-FFA), Artificial Neural Network (ANN), and Support Vector Machine–Wavelet (SVM-Wavelet) have emerged as powerful computational models. These have also shown significant success in dealing with massive data analysis for a large number of applications.

This book, *Advanced Soft Computing Techniques in Data Sciences, IoT and Cloud Computing*, aims to play a significant role in improvising human life to a great extent. All the researchers and practitioners who are working in the field of computer engineering, medicine, biology application, signal processing, and mechanical engineering will be highly benefited.

This book would be a good collection of state-of-the-art approaches for soft computing-based applications to various engineering fields. It will be very beneficial for the new researchers and practitioners working in the field to quickly know the best-performing methods. They would be able to compare different approaches and can carry forward their research in the most important area of research which has a direct impact on the betterment of the human life and health. This book would be very useful because there is no book in the market that provides a good collection of state-of-the-art methods of soft computing-based models for Multimedia retrieval, Fault tolerance, System modeling, Network architecture, Web semantics, big data analytics, time series, and biomedical and health informatics.

Objective

The purpose of this book is to report the latest advances and developments in the field of Multimedia retrieval, Fault tolerance, System modeling, Network architecture, Web semantics, big data analytics, time series, and biomedical and health informatics. The book comprises the following three parts:

- Soft Computing Techniques for Internet of Things (IoT) devices
- Soft Computing Techniques in Cloud Computing and Computer Networking
- Soft Computing Techniques in Data Science

Organization

The book, "Advanced Soft Computing Techniques in Data Sciences, IoT and Cloud Computing", consists of 19 edited chapters and the whole contents of the book are organized into the following three sections:

- *Part 1: Soft Computing Techniques for IoT Devices.* This section has focused on Soft Computing Techniques for Internet of Things paradigms and their application in wearable assistive devices for visually impaired people, security of IoT devices, medical data security in healthcare systems, content-based video retrieval systems, and also applications in intelligent farming. There are six chapters in this section. The first chapter looks into the application of soft computing techniques for designing IoT-based wearable assistive devices for visually impaired people. The second, third, fourth, and fifth contributions focus on the security of IoT devices and their application in medical data in healthcare systems and content-based video retrieval systems. The sixth chapter discusses various existing soft computing techniques and their applications for decision support in intelligent farming.

- *Part II: Soft Computing Techniques in Cloud Computing and Computer Networking.* The second part comprises five chapters. The first contribution discusses about the hybrid cloud data protection using the machine learning approach. The second chapter audits the analysis of LSTM networks for COVID-19 impact prediction. The third chapter focuses on the progress and systematic review of soft computing techniques for energy consumption and resource allocation on the cloud. The fourth chapter discusses automatic segmentation and classification of Brain Tumor from MR Images using DWT-RBFNN.
- The fifth chapter discusses on the application of unsupervised learning on automatic localization of optic disc in Retinal Fundus Image.
- *Part III: Soft Computing Techniques in Data Science.* There are eight chapters in this section. The first chapter discusses the performance evaluation of Hybrid Machine Learning Algorithms for Medical Image Classification. The second chapter discusses the issues and challenges of soft computing techniques for Healthcare Decision Support Systems and the third chapter focuses on the truth values of Modus Ponens and Modus Tollens Rule for Linguistic Truth-Valued Propositions and Its Application in Taking Decisions in Health Care. The fourth chapter focuses on a case study of Amazon based on the analysis of customers' reviews using Soft Computing Classification Algorithms. The fifth chapter throws light on Pattern Mining—FTISPAM applying Hybrid Genetic Algorithm and the sixth chapter highlights soft computing techniques for Medical Diagnosis, Prognosis, and Treatment. The concluding chapter explains Artificial Intelligence Applications for COVID-19 Pandemic and the eighth chapter deliberates on the prediction of transmittable diseases in a location using ARIMA.

Target Audiences

The current volume is a reference text aimed to support a number of potential audiences, including the following:

- Researchers in this field who wish to have the up to date knowledge of the current practice, mechanisms, and research developments.
- Students and academicians of the biomedical and informatics field who have an interest in further enhancing the knowledge of the current developments.

- Industry and people from Technical Institutes, R & D Organizations and working in the field of machine learning, IoT, Cloud Computing, biomedical engineering, health informatics, and related fields.

Baripada, India — Sujata Dash
Bhubaneswar, India — Subhendu Kumar Pani
Auburn, USA — Ajith Abraham
Baltimore, USA — Yulan Liang

Acknowledgements

The editors would like to acknowledge the help of all the people involved in this project and, more specifically, the reviewers who took part in the review process. Without their support, this book would not have become a reality.

First, the editors would like to thank each one of the authors for their time, contribution, and understanding during the preparation of the book.

Second, the editors wish to acknowledge the valuable contributions of the reviewers regarding the improvement of quality, coherence, and content presentation of chapters.

Last but not the least, the editors wish to acknowledge the love, understanding, and support of their family members during the preparation of the book.

Odisha, India — Sujata Dash
Bhubaneswar, India — Subhendu Kumar Pani
Auburn, USA — Ajith Abraham
Baltimore, USA — Yulan Liang

Contents

Soft Computing Techniques for IoT Devices

A Wearable Assistive Device for Safe Travel Using Transfer Learning and IoT for Visually Impaired People 3
R. Priyatharshini, R. Senthil Kumar, M. Sanjay Sivakumar, A. Mathumathi, and Nancy Sharon Johnson

Soft Computing Techniques for Physical Layer Security of IoT Devices 27
C. Ismayil Siyad and S. Tamilselvan

Linear Congruence Generator and Chaos Based Encryption Key Generation for Medical Data Security in IoT Based Health Care System 53
Anirban Bhowmik, Sunil Karforma, Joydeep Dey, and Arindam Sarkar

Content Based Video Retrieval—Methods, Techniques and Applications 81
Reddy Mounika Bommisetty, P. Palanisamy, and Ashish Khare

Building the World of Internet of Things 101
Seema Sahai, Richa Goel, and Gurinder Singh

Applicability of Machine Learning Algorithms for Intelligent Farming 121
Bharti Verma, Nikhil Sharma, Ila Kaushik, and Bharat Bhushan

Soft Computing Techniques in Cloud Computing and Computer Networking

Hybrid Cloud Data Protection Using Machine Learning Approach 151
D. Praveena, S. Thanga Ramya, V. P. Gladis Pushparathi, Pratap Bethi, and S. Poopandian

Analysis of Long Short Term Memory (LSTM) Networks in the Stateful and Stateless Mode for COVID-19 Impact Prediction 167
Vinayak Ashok Bharadi and Sujata S. Alegavi

Soft Computing Techniques for Energy Consumption and Resource Aware Allocation on Cloud: A Progress and Systematic Review 191
Sukhpreet Kaur, Yogesh Kumar, and Sushil Kumar

Automatic Segmentation and Classification of Brain Tumor from MR Images Using DWT-RBFNN 215
Hari Mohan Rai, Kalyan Chatterjee, and Anand Nayyar

Automatic Localization of Optic Disc in Retinal Fundus Image Based on Unsupervised Learning 245
J. Prakash and B. Vinoth Kumar

Soft Computing Techniques in Data Science

Performance Evaluation of Hybrid Machine Learning Algorithms for Medical Image Classification 281
N. T. Renukadevi

Computing Truth Values of Modus Ponens and Modus Tollens Rule for Linguistic Truth-Valued Propositions and Its Application in Taking Decisions in Health Care 301
Sumita Basu and Bithi Chattaraj

Analysis of Customers' Reviews Using Soft Computing Classification Algorithms: A Case Study of Amazon 331
Koushal Kumar and Bhagwati Prasad Pande

Pattern Mining—FTISPAM Using Hybrid Genetic Algorithm 353
L. Mary Gladence, S. Shanmuga Priya, A. Shane Sam, Gladis Pushparathi, and E. Brumancia

Soft Computing Techniques for Medical Diagnosis, Prognosis and Treatment 371
Surabhi Adhikari, Surendrabikram Thapa, and Awishkar Ghimire

Role of Artificial Intelligence in COVID-19 Pandemic 401
Abhishek Mehta and Trupti Rathod

Prediction of Transmittable Diseases Rate in a Location Using ARIMA 415
Varun Totakura and E. Madhusudhana Reddy

Editors and Contributors

About the Editors

Sujata Dash is currently working as an Associate Professor of Computer Science at North Orissa University in the Department of Computer Science, Baripada, India. She has been consistently delivering her services in teaching and guiding students for more than two and a half decades now. She is a recipient of Titular Fellowship from the Association of Commonwealth Universities, UK. She has worked as a visiting professor of the Computer Science Department of the University of Manitoba, Canada. She has published more than 170 technical papers in international journals/ proceedings of international conferences/ edited book chapters of reputed publishers like Springer, Elsevier, IEEE, IGI Global USA, 10 national and international patents, and has published many textbooks, monographs, and edited books. She is a member of international professional associations like ACM, IRSS, CSI, IMS, OITS, OMS, IACSIT, IST, IEEE and is a reviewer of around 15 international journals which include World Scientific, Bioinformatics, Springer, IEEE ACCESS, Inderscience, and Science Direct publications. In addition, she is a member of the editorial board of around 10 international journals. She has visited many countries and delivered keynotes, invited speech, and chaired many special sessions in International conferences in India and abroad. Her current research interest includes Machine Learning, Data Mining, Big Data Analytics, Bioinformatics, Soft Computing, and Intelligent Agents.

Subhendu Kumar Pani received his Ph.D. from Utkal University Odisha, India in the year 2013. He is working as a Professor in the Department of Computer Science and Engineering and also a Research coordinator at Orissa Engineering College (OEC) Bhubaneswar. He has more than 16 years of teaching and research experience. His research interests include Data mining, Big Data Analysis, web data analytics, Fuzzy Decision Making, and Computational Intelligence. He is the recipient of 5 researcher awards. In addition to research, he has guided two Ph.D. students and 31 M.Tech students. He has published 7 patents and 51 International Journal papers (25

Scopus index). His professional activities include roles as Associate Editor, Editorial board member, and/or reviewer of various International Journals. He serves as Associate Editor/Editorial Board Member for more than 50+ International Journals/Conferences of high repute and impact. He is an Associate with no. of conference societies. He has more than 100 international publications, 5-authored books, 4 edited books, and 10 book chapters into his account. He is a fellow in SSARSC and the life member in IE, ISTE, ISCA, OBA, OMS, SMIACSIT, SMUACEE, and CSI.

Ajith Abraham is current working as the Director of Machine Intelligence Research Labs (MIR Labs), which has members from more than 100 countries. Dr. Abraham's research and development experience includes more than 27 years in the industry and academia. He received the M.S. degree from the Nanyang Technological University (NTU), Singapore, and the Ph.D. degree in Computer Science from Monash University, Melbourne, Australia. He worked in a multidisciplinary environment involving machine (network) intelligence, cybersecurity, sensor networks, Web intelligence, scheduling, and data mining and applied to various real-world problems. He has given more than 100+ conference plenary lectures/tutorials and invited seminars/lectures in over 100 Universities around the globe.

Yulan Liang holds a Ph.D. in Statistics and has been a tenured faculty at the University of Maryland, Baltimore (UMB) since 2008. Before that, she worked as a tenure-track assistant professor at the School of Public Health and School of Medicine at the University at Buffalo. Dr. Liang's expertise is in the cutting-edge multidisciplinary fields including big data science, statistics, data mining, neural networks, machine learning, Bayesian inferences, artificial intelligence, and bioinformatics with applications to large biomedical, health science, omics, and healthcare data for risk assessments, predictions, and medical decision-making. She has published more than a hundred peer-reviewed journal and conference proceeding articles in the data science fields. She was funded as PI by NSF and served as a statistics core director of NIH P30, U01, and P20 grants and co-Investigator on multiple NIH-funded grants. She also served as co-Investigator and biostatistician for private organizations in health associations and other funding agencies, including the Robert Wood Johnson Foundation, Alzheimer's Association, and National Council of State Boards of Nursing.

Contributors

Surabhi Adhikari Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

Sujata S. Alegavi Department of Electronics Engineering, Thakur College of Engineering and Technology, Mumbai, India

Sumita Basu Department of Mathematics, Bethune College, University of Calcutta, Kolkata, West Bengal, India;
Guest Faculty, Department of Mathematics, The Heritage College, University of Calcutta, Kolkata, West Bengal, India

Pratap Bethi OPG Power Generation Pvt. Ltd., Chennai, India

Vinayak Ashok Bharadi Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India

Anirban Bhowmik Department of Computer Science, M.U.C. Women's College, Burdwan, WB, India

Bharat Bhushan School of Engineering and Technology, Sharda University, Greater Noida, India

Reddy Mounika Bommisetty Department of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, Tiruchirappalli, India

E. Brumancia Sathyabama Institute of Science and Technology, Chennai, India

Bithi Chattaraj Department of Mathematics, Bethune College, University of Calcutta, Kolkata, West Bengal, India

Kalyan Chatterjee Department of Electrical Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India

Joydeep Dey Department of Computer Science, M.U.C. Women's College, Burdwan, WB, India

Awishkar Ghimire Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

V. P. Gladis Pushparathi Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai, India

Richa Goel Amity International Business School, Amity University Uttar Pradesh, Noida, India

C. Ismayil Siyad Pondicherry Engineering College, Puducherry, India

Nancy Sharon Johnson Easwari Engineering College, Chennai, India

Sunil Karforma Department of Computer Science, The University of Burdwan, Burdwan, WB, India

Sukhpreet Kaur Department of Computer Science & Engineering, Chandigarh Engineering College Landran, Mohali, India

Ila Kaushik Krishna Institute of Engineering & Technology, Ghaziabad, India

Ashish Khare Department of Electronics and Communication Engineering, University of Allahabad, Prayagraj, India

Koushal Kumar Sikh National College, Qadian, Guru Nanak Dev University, Amritsar, Punjab, India

Sushil Kumar Chandigarh Engineering College, Landran, India

Yogesh Kumar Department of Computer Science & Engineering, Chandigarh Group of Colleges, Mohali, India

E. Madhusudhana Reddy Department of Computer Science and Engineering, Guru Nanak Institutions Technical Campus, Hyderabad, India

L. Mary Gladence Sathyabama Institute of Science and Technology, Chennai, India

A. Mathumathi Easwari Engineering College, Chennai, India

Abhishek Mehta Faculty of IT and Computer Science, Parul University, Vadodara, Gujarat, India

Anand Nayyar Graduate School, Duy Tan University, Da Nang, Vietnam;
Faculty of Information Technology, Duy Tan University, Da Nang, Vietnam

P. Palanisamy Department of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, Tiruchirappalli, India

Bhagwati Prasad Pande Department of Computer Applications, LSM Government PG College, Pithoragarh, Uttarakhand, India

S. Poopandian Infosys Pvt., Ltd., Chennai, India

J. Prakash Department of CSE, PSG College of Technology, Coimbatore, India

D. Praveena Department of Information Technology, R.M.D Engineering College, Kavaraipettai, India

R. Priyatharshini Easwari Engineering College, Chennai, India

Gladis Pushparathi Velammal Institute of Technology, Chennai, India

Hari Mohan Rai Department of Electronics and Communication Engineering, Krishna Engineering College, Ghaziabad, UP, India

Trupti Rathod Vidyabharti Trust College of MCA, Gujarat Technological University, Bardoli, Gujarat, India

N. T. Renukadevi Department of CT-UG, Kongu Engineering College, Erode, India

Seema Sahai Amity International Business School, Amity University Uttar Pradesh, Noida, India

M. Sanjay Sivakumar Easwari Engineering College, Chennai, India

Arindam Sarkar Department of Computer Science and Electronics, R.K.M. Vidyamandira, Belur Math, WB, India

R. Senthil Kumar Easwari Engineering College, Chennai, India

A. Shane Sam St. Joseph's College of Engineering, Chennai, India

S. Shanmuga Priya M.I.E.T Engineering College, Trichy, India

Nikhil Sharma Department of Computer Science & Engineering, Delhi Technological University, Delhi, India

Gurinder Singh Amity International Business School, Amity University Uttar Pradesh, Noida, India

S. Tamilselvan Pondicherry Engineering College, Puducherry, India

S. Thanga Ramya Department of Information Technology, R.M.D Engineering College, Kavaraipettai, India

Surendrabikram Thapa Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

Varun Totakura Tata Consultancy Services, Hyderabad, Telangana, India

Bharti Verma HMR Institute of Technology & Management, Delhi, India

B. Vinoth Kumar Department of IT, PSG College of Technology, Coimbatore, India

Abbreviations

2D/3D	Two Dimension/Three Dimension
3D	Three-dimensional
AANN	Adaptive Artificial Neural Network
ABC	Artificial Bee Colony
ACO	Ant Colony Optimization
AD	Alzheimer's Disease
AdaBoost	Adaptive Boost
ADNI	Alzheimer's Disease Neuroimaging Initiative
AES	Advanced Encryption Standard
AFBNN	Adaptive Firefly Backpropagation Neural Network
AFCNN	Advanced Fuzzy Cellular Neural Network
AI	Artificial Intelligence
AIC	Akaike's Information Criterion
ANFIS	Adaptive Neuro Fuzzy Inference System
ANFIS	Adaptive Neuro-fuzzy Inference System
ANN	Artificial Neural Network
ANN	Artificial Neural Networks
API	Application Programming Interface
AR	Auto Regressive
ARIMA	Auto Regressive Integrated Moving Averages
ARMA	Auto Regressive Moving Averages
ARTISANA	Artificial Intelligence in Sleep Analysis Algorithm
ASIN	Amazon Standard Identification Number
ATD	Artificial Training Data
BA	Bat Algorithm
BFO	Bacterial Foraging Optimization
BLE	Bluetooth Low Energy
BMRI	Brain Magnetic Resonance Imaging
BP	Back Propagation
BPNN	Back Propagation Neural Network
BRIEF	Binary Robust Independent Elementary Features

BRIoT	Behavior Rule specification-based misbehavior detection for IoT
BRISK	Binary Robust Invariant Scalable Key point
C	Cases
CAD	Computer-Aided Diagnosis
CANFIS	Coactive Neuro-Fuzzy Inference System
CBDHDSE	Cross-Border Digital Healthcare Decision Support Ecosystem
CBIR	Content-Based Image Retrieval
CBSE	Component-Based Software Engineering
CBVR	Content-Based Video Retrieval
CDRC	Consumer Data Research Centre
CE	Convergent Encryption
CFT	Clustering Fusion Technique
CGS	Chronic-disease Guideline Support
CG-SVM	Coarse Gaussian SVM
CIR	Channel Impulse Response
CK	Component Kosh
CLARA	Clustering Large Applications
CloudSim	Cloud Simulator
CM	Classification Methods
CMS	Content Management System
CNN	Convolution Neural Network
CNN	Convolutional Neural Networks
CNTK	Cognitive Toolkit
COVID-19	Novel Coronavirus Disease
CPNN	Convolutional Pre-processing Neural Network
CPU	Central Processing Unit
CR	Coverage Ratio
CRNN	Convolutional Recurrent Neural Network
CSI	Channel State Information
CSP	Cloud Service Provider
CT	Computed Tomography
CW	Continuous Wavelet
DAM	Data-Adaptive Matrix
DAS	Distance between Adjacent Signals
DBN	Deep Belief Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DDoS	Distributed Denial of Attack
DDPA	Deduplication Processing Algorithm
DFM	Data Fusion of MRI
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DOI	Document Identifier
DS	Dataset Size
DSA	Digital Signature Algorithm
DSRBACA	Dynamic Spatial Role-Based Access Control Algorithm

DT	Decision Tree
DT	Decision Trees
DW-LSTM	Dynamic Watermarking LSTM
DWT-RBFNN	Discrete Wavelet Transform–Radial Basis Function Neural Network
ECDH	Elliptic Curve Diffie–Hellman
ECG	Electrocardiogram
ECG	Electrocardiography
EHR	Electronic Health Record
EL	Ensemble Learning
ELM-IPSO	Extreme learning machine-improved particle swarm optimization
EMG	Electromyography
ERDIP	Electronic Record Development and Implementation Program
FA	False Alarm
FAST	Features from Accelerated Segment Test
FCFS	First Come First Serve
FCM	Fuzzy c-means
FCM	Fuzzy C Means
FCM	Fuzzy C-Mean
FCMC	Fuzzy C-Means Clustering, MLS
FDR	False Discovery Rate
FFNN	Feed Forward Neural Networks
FIVR	Fine-grained Incident Video Retrieval
FKM	Fuzzy K Means
FN	False Negative
FP	False Positive
FS	Feature Selection
FS	Fuzzy Systems
FT	Fourier Transform
FTISPAM	Fuzzy Time Interval Sequential Pattern Mining
FWT	Fast Wavelet Transform
GA	Genetic Algorithm
GA-ANN	Genetic Algorithm–Artificial Neural Network
GBHSP	Gene-Based Health Screening Personalization
GLCM	Gray Level Co-occurrence Matrix
GLOGTH	Global and Local Oriented Gabor Texture Histogram
GOA	Grasshopper Optimization
GPU	Graphical Processing Units
HBR	Harvard Brain Repository
HCAPN	Hierarchical Context-Aware Aspect-Oriented Petri Net
HGA	Hybrid Genetic Algorithm
HIPAA	Health Insurance Portability and Accountability Act
HOG	Histogram of Oriented Gradients
HOG	Histogram of the oriented gradient
HRF	High-Resolution Fundus

HSO	Hybrid Swarm Optimization Algorithm
HSP	Health Screening Personalization
HSV	Hue, Saturation, Value
I	Images
I/O	Input/Output
IaaS	Infrastructure as a Service
ICA	Independent Component Analysis
ICT	Information Communication Technology
IDF	Inverse Document Frequency
IIoT	Industrial IoT
IoT	Internet of Things
IoTD	IoT Devices
KM	Kernel Machine
KMC	K-Means Clustering
KNN	K-Nearest Neighbor
LBP	Local Binary Pattern
LBPV	LBP Variance
LC	Linear Classifiers
LDPC	Low Density Parity Check
LIA	Lattice Implication Algebra
LOG	Laplacian of Gaussian
LOS	Line of Sight
LPLA	Lightweight Physical Layer Authentication
LSTM	Long Short Term Memory
LSTM	Long Short Term Memory Networks
LSTM	Long Short-Term Memory
LT	Luby Transform
LTVP	Linguistic Truth-Valued Propositions
MA	Moving Average
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAFD	Modify Adaptive Frame Differencing
MAP	Mean Arterial Pressure
mAP	Mean Average Precision
MATLAB	Matrix Laboratory
MCI	Most Common Imputation
MD	Multipath Delay
MDPSO	Modified Discrete Particle Swarm Optimization
MEI	Mean Imputation
MI	Moment Invariant
ML	Machine Learning
MLP	Multilayer Perceptron
MLP	Multi-layer Perceptron
MOGA	Multi-Objective Genetic Algorithm
MR	Magnetic Resonance

MRI	Magnetic Resonance Imaging
MRS	Magnetic Resonance Spectroscopy
MSE	Mean Square Error
MSE	Mean Squared Error
MSER	Maximally Stable Extremal Regions
MSER	Maximum Stable Extremal Regions
MSNE	Mixed Strategy Nash Equilibrium
NBC	Naïve Bayes Classifiers
NBTS	National Brain Tumor Society
NER	Named-Entity Recognition
NFC	Near Field Communication
NLOS	Non-Line of Sight
NLP	Natural Language Processing
NR	Novelty Ratio
NTM	New Thresholding Method
OASIS	Open Access Series of Imaging Studies
OCC	One Class Classification
OCR	Optical Character Recognition
OS	Operating System
OSCR	Ontology of Software Component for Retrieval
OWL	Web Ontology Language
P	Patients
PA	Physician’s Assistant
PACS	Picture Archiving and Communication Systems
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PD	Proton Density
PEAM	Performance Evaluation Assessment Metrics
PET	Positron Emission Tomography
PID	Pima Indian Diabetes
PLS	Physical Layer Security
PM	Practice Management
PN	Pseudo Noise
POS	Part of Speech
PPA	Peripapillary Atrophy
PPV	Positive Predicted Value
PSA	Particle Swarm Algorithm
PSD	Power Spectral Density
PSO	Particle Swarm Optimization
QLIA	Quasi-Lattice Implication Algebra
QLTVP	Quasi-Linguistic Truth-Valued Propositions
QoS	Quality of Service
RACS	Role-Based Access Control Scheme
RAI	RAdiometric Identification
RBAC	Role-Based Access Control

RBF	Radial Basis Function
RBM	Restricted Boltzmann Machines
RCM	Revenue Cycle Management
RCT	Randomized Control Trials
RCUO	Retrieval of the Component Using Ontology
ReLU	Rectified Linear Unit
RF	Random Forest
RFID	Radio Frequency Identification
RGB	Red, Green, and Blue
RGM	Region Growing Method
RIM	Reference Information Model
RMS	Root Mean Square
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
ROI	Region of Interest
RP-CNN	CNN and Recurrence Plot
RR	Round Robin Algorithm
RS	Reed Solomon
RSA	Rivest Shamir Adleman
RSS	Received Signal Strength
SC	Sentiment Classification
SC	Soft Computing
SCI	Similarity Computation Index
SD	Standard Deviation
SGLD	Spatial Gray Level Dependence Matrix
SHA	Secure Hash Algorithm
SI	Swarm Intelligence
SIFT	Scale Invariant Feature Transform
SLA	Service Level Agreement
SM	Segmentation Methods
SOA	Service-Oriented Architecture
SOAP	Simple Object Access Protocol
SOM	Self-Organizing Map (SOM)
SPAM	Sequential Pattern Mining
SPARQL	SPARQL Protocol and RDF Query Language
SPDTH	Similarity-Preserving Deep Temporal Hashing
SPECT	Single Photon Emission Computed Tomography
SPECT	Single Proton Emission Computed Tomography
SR	Sparse Representation
STFT	Short-Time Fourier Transform
SURF	Speeded Up Robust Feature
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency–Inverse Document Frequency
TLBO	Teaching Learning-Based Optimization

TN	True Negative
TP	True Positive
TPU	Tensor Processing Units
U.S	United States
UI	User Interface
UK	United Kingdom
ULBP	Uniform Local Binary Pattern
URL	Uniform Resource Locator
US	Ultrasound
USRP	Universal Software Radio Peripheral
VOC	Pascal Visual Object Classes
WHO	World Health Organization
WLBP	Weber Local Binary Pattern
YOLO	You Only Look Once

Soft Computing Techniques for IoT Devices

A Wearable Assistive Device for Safe Travel Using Transfer Learning and IoT for Visually Impaired People

R. Priyatharshini, R. Senthil Kumar, M. Sanjay Sivakumar, A. Mathumathi, and Nancy Sharon Johnson

Abstract There are a number of discomforts faced by visually impaired people every day in both indoor and outdoor surroundings. Assistive Technology for people with visual disabilities plays a vital role in their Independent living. Various systems have been developed to help them to live a better life even with the low or no vision. Visual Mobility plays a vital role in their Independent living such as arrival of buses, recognizing the Route Number from number plates, finding the doorstep in a train etc. This paper proposes an efficient approach for recognizing the Route details from a bus that helps in easy commuting in bus stations for the visually impaired. The You Only Look Once (YOLO v3) model can be used to detect the real time object bus and also segment the bus name board i.e. the Region of Interest (ROI) by transfer learning it with a custom dataset. To further improve the pattern recognition accuracy the numbers of anchoring boxes were increased from 3 to 5 in all the strides which provided more precise results. The route details present in the segmented ROI is converted into text using Tesseract tool that uses LSTM (Long Short Term Memory) engine for producing text from recognized characters in image and then regular expression is used to filter out only the bus number (alphanumeric or numeric values). The obtained bus number is converted as voice output (using text to speech library e-speak) along with bus details based on the bus number extracted from the number plate. An approach to detect the door step at railway compartments has also been proposed for visual mobility. During the travel, the device can be further used to know the current location and the distance remaining to reach the destination.

Keywords Pattern recognition · Deep learning · Image processing · Transfer learning

1 Introduction

Globally the number of humans of all ages who are visually impaired is estimated to be around 285 million. Partial or complete loss of vision is common among most of

R. Priyatharshini (✉) · R. Senthil Kumar · M. Sanjay Sivakumar · A. Mathumathi · N. S. Johnson
Easwari Engineering College, Chennai, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_1

the old age people and it is difficult for a person to deal with such kind of impairment after experiencing niche lifestyle with perfect vision. Visually impaired people has less ability to perform everyday task like other people so their quality of life gets affected and lose the ability to experience the necessary things in the world. Visually impaired people are comfortable only with their familiar places because they know the layout of the place and they would have memorized it, but when it comes to new environment they may find several difficulties. There are a number of technological advancements and everyday developments which are focused on providing a better lifestyle for handicapped and old age peoples. Although there are bright ideas and thoughts, there are practical difficulties in implementing them as a product that can be used every day by this section of the population.

One of the leading difficulties for visually impaired people in day to day scenarios is to travel alone without any dependencies from others. Some of them can have the blurry scene of a bus arriving at the bus terminal, still recognizing the bus route remains a challenging and an impossible task for them. People also feel embarrassed to request others to help them to identify bus numbers. Another major problem they face is in the railway stations. It is difficult for visually impaired person to choose which exit side of the train to use to get out from the train. Sometimes they accidentally trip over the wrong exit side and fall down to the track.

These types of incidents has occurred several times. These all leads to rise in demand to provide a simple and powerful solution to address these problems. Most of the time they are present alone in the station and they find it difficult to receive any assistance from other people. Numbers of bus recognition system ideas are introduced, like identifying the bus route with help of smartphone using Maximum Stable Extremal Regions (MSER) applications, or in another system histogram of the oriented gradient (HOG) identifies the codes of texts from detected text regions and produces output as audio notification. There are also systems introduced for bus detection using wireless networks and satellite signals. It requires installation of sensors and other modules in bus and bus stations along with periodic maintenance.

Most of these current systems lack performance required for real time detection and usage along with few other technical snags. There are a number of breakthrough researches and technologies coming out throughout the year. They highly involve attributes like quick performance, efficient methodologies, less implementation cost and many others. Therefore, coining enhanced solutions to all kinds of problems is not difficult and this includes for the above stated problems of visually impaired too.

2 Review of Literature

In the research field of computer vision and in other areas a number of bus recognition systems are proposed; however, most of them use active devices and sensors such as Global Positioning System tracking system, RFID (Radio Frequency Identification), Beacons etc. For vision based approaches, Wongta [1] introduced a system that uses MSER (Maximally Stable Extremal Regions) to recognize bus numbers. Their solution finds all the texts in an image rather than the required ones i.e. the bus number. Guida et al. [2] proposed a system for bus route number which uses a number of interlinked classifiers along with adaboost in order to identify the route number or other elements present in the front of the bus and some corrections are done to the extracted features to obtain rectified numeric values. The object is then converted to Hue, Saturation, Value (HSV) colorspace and then partitioning of each numeric or digit value. Finally Optical character Recognition (OCR) is used to identify the digits and output is produced as voice to the user. Viola and Jones [3] proposed a face detection framework that involved three key contributions which were integral images, ada boost and cascading classifiers. It was a simple and powerful technique to identify faces from binary images. Pan et al. [4] proposed a bus detection system to help the visually impaired where Histogram of the oriented gradients is used to extract features from bus images and a cascading Support Vector Machine model is implemented for a bus classifier to detect the bus facade in frames of windows. Tsai and Yeh [5] also introduced a system for bus detection to support the blind. The system included the functionalities of detection of the moving bus, bus panel detection and text detection from the text region of the panel. The system made use of MAFD(Modify Adaptive Frame Differencing) and was found to have great accuracy when features were extracted from different frames of the video. In [6] Bouhmed used an ultrasonic sensor and camera in a walking cane to identify obstacles in the path and inform it to the blind. The system produces output through voice. Zahir et al. [7] developed a prototype of wearable head mountable device by customizing the Virtual Reality glass with ultrasonic sensors and HC-SR04 because it take minimum amount of time for detection and can also find obstacle in the longer range. The prototype is developed using Arduino. Ani et al. [8] introduced a voice assistive text reading system that is integrated with eyeglasses. A Camera is inbuilt with the eyeglass to capture a image and with Tesseract-OCR text is extracted from the captured image. Open Software E-speak is used for TTS. A method for end-to-end real-time scene text localization and recognition is presented in [9] and the robustness of the proposed method against noise and low contrast of characters is well demonstrated using “false positives”. In [12] a Convolutional Neural Networks (CNN) model is developed to detect English and Thai text from natural scene images with improved accuracy compared to the other existing methods A method for Multi-script Text Extraction from Natural Scenes is presented in [10] which exploits collaboration of proximity and similarity laws to create text-group hypotheses. The problem of establishing correspondences between a pair of images taken from different viewpoints is analysed in [11]. There are a number of techniques and various innovative ideas proposed to help the blind. Most of these systems are dynamic and they have less difficulty when it comes to real time usage. Our proposed system is for real time usage and uses video recognition rather than capturing only images. The bus arrival and waiting time are always varying therefore every frame must be checked instead of single image snaps which would be less efficient for the detection of bus and bus board. The system proposed can be integrated with any of the existing ones, where the components are almost similar or it can be developed as a separate one in which any other new features can be extended.

3 Overview of Proposed System

A visually impaired person after loss of vision limits them from indulging in any activities that involves travelling in public transport facilities. They become hesitant to ask help from other people. For them to able to live their life independently we have developed Assistive Device for visually impaired to people that helps them to travel.

A single integrated system is developed to assist the visually impaired. User can choose the required operation by commanding through voice. Raspberry Pi serves the processing unit for all the operations and it is connected to an external portable power supply. An approach to accurately detect the real time bus object and to segment the ROI for extracting and recognizing the route details has been proposed to assist visually impaired people. An enhanced version of the You Only Look Once algorithm is proposed in this paper which detects the objects in an image at a single instance. It is an algorithm based on regression and it was pre trained on a number of classes.

The algorithm initially splits the image into cells of 19 × 19 grid and each cell is responsible for predicting 5 bounding boxes. Each of the bounding boxes can be described using 5 descriptors as shown in (Fig. 1). They are,

1. Width ($\mathbf{b_w}$): Width of the bounding box.
2. Height ($\mathbf{b_h}$): Height of the bounding box.
3. Center of a bounding box ($\mathbf{b_x}$ $\mathbf{b_y}$): Center coordinates of the bounding box.
4. Value **c**: Corresponding to a class of an object (i.e. car, traffic lights)
5. Value $\mathbf{p_c}$: Probability that there is an object in the bounding box.

For areas under each of the bounding box, the Convolution Neural Network (CNN) processes to identify to which class the object under the area belongs. The boxes with low object probability (p_c) and unknown classes (**c**) are removed. Areas under high overlapping boundary regions across cells along with greater p_c are the detected objects in the image. By transfer learning the YOLO v3 model with a custom dataset, the same above procedure is involved to detect the bus board in our system. The detected bus board image is then used to obtain the text with help of tesseract optical character recognition tool. It gathers the outlines into blobs by nesting. Blobs are further ordered into lines of text and each of these text lines are drilled down into probable terms and further into characters. Multiple passes are involved before finally predicting the character, words and the sentences. Latest versions of tesseract use deep learning frameworks which are accurate and fast. Output from the OCR is filtered with regular expressions which are constrained to find at maximum three characters length data that are either numeric or alphanumeric. The obtained data is the bus number as bus route number usually falls under these constraints in modern bus route naming systems. The identified bus number is informed to blind through text to speech method (Fig. 2).

Speaker/headphones are present to give output notifications. Microphone is used to obtain the voice input from the user. If the voice input is found to be difficult for

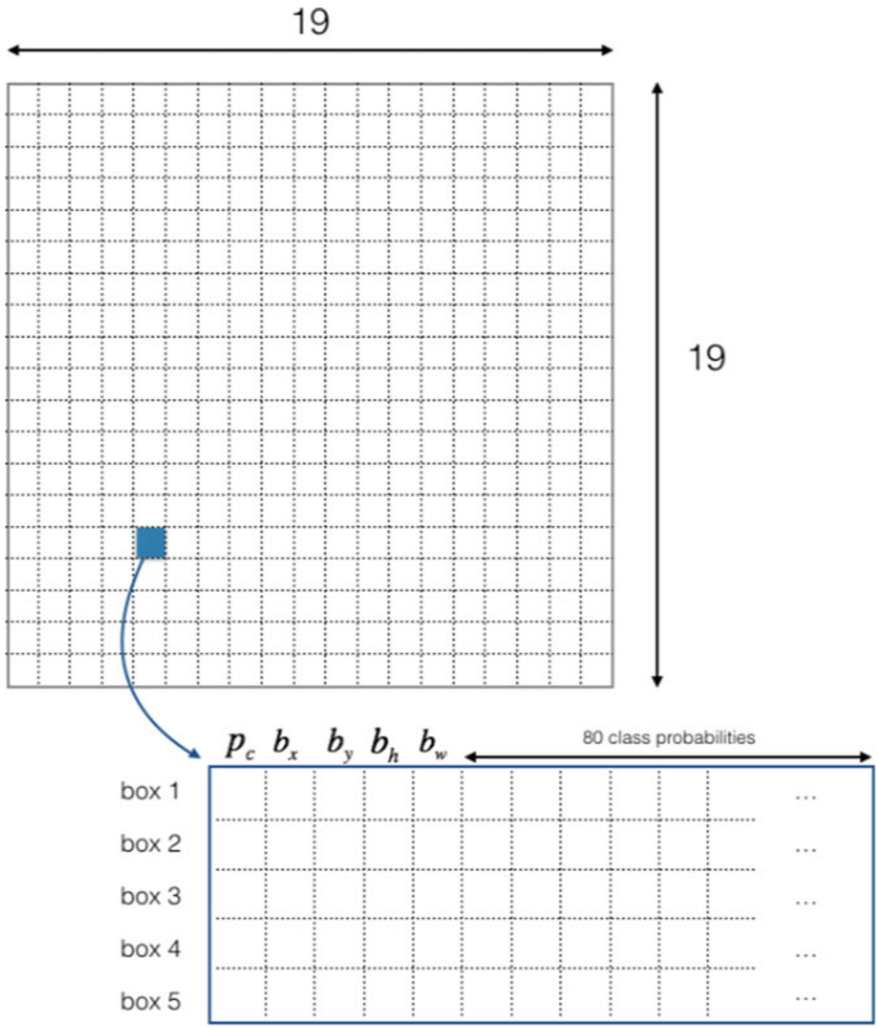
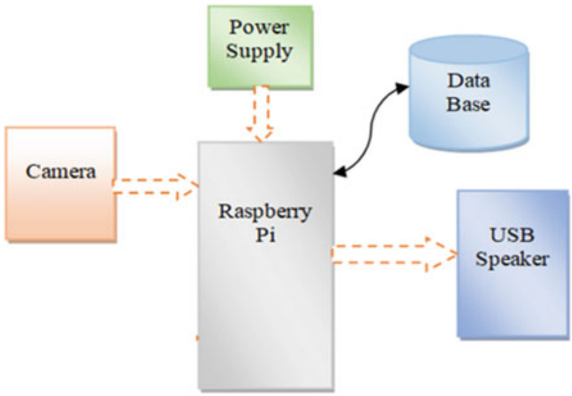


Fig. 1 A 19 × 19 grid with descriptors (*Source* https://pylessons.com)

the user or if there is a lot of noise, button inputs can be added to the system. GPS can be added as an extended feature which can inform the user where they are currently located and to know the remaining time it would take to reach the destination.

For railway platform detection, an Ultrasonic Sensor is present at a certain angle facing ground every time. It is also connected to the raspberry pi and it gives the frequent depth measures from ground to the pi in which the depth identification algorithm is executed. Speaker/headphones are present to give output notifications. Microphone is used to obtain the voice input from the user. If the voice input is found to be difficult for the user or if there is lot of noise, button inputs can be added to the

Fig. 2 Proposed system components architecture



system. GPS can be added as an extended feature which can inform the user where they are currently located and to know the remaining time it would take to reach the destination.

4 Proposed Methodology

The Proposed system mainly consists of two sub modules that are chosen for different scenarios by the user. A methodology to detect the real time object (bus) and segment the number plate (region of interest) has been proposed for recognizing the route number in a bus. In this approach, there are three sub modules involved. First the bus object is recognized and the bus board is segmented for which the core functionality is based on Convolution Neural Network and transfer learning. For the next module, tesseract optical character recognition engine is used to recognize the text and finally with regular expressions, the bus route number is recognized (Fig. 3).

4.1 Bus Board Detection for Route Information Extraction

This subsystem deals with identifying the bus board in real time. A high definition wide angle camera is mounted to the head wear and is connected to the pi. The camera is used to capture the bus arrival scene and from the video frames, ROI detection and extraction takes place.

More than 400 images were used to create a customized model. Sample images from training dataset is shown in Fig. 4. The images obtained were in a daylight environment. Preprocessing is done to remove the blurry and dark images before initiating the training process. The region of interests were marked and their labeling was given. Pascal visual object classes (VOC) is a format for providing object detection data, i.e.

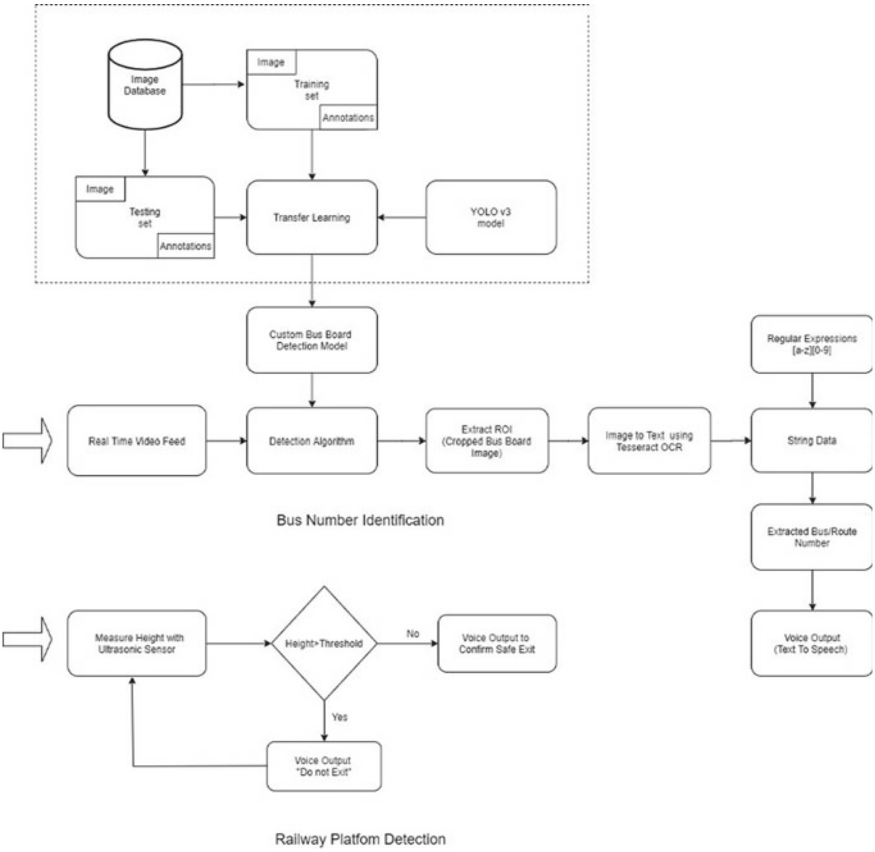


Fig. 3 Overview of proposed system

images with bounding boxes. Using the annotation tool, the bus boards are marked with bounding boxes in each image manually and these values are saved in a xml file. The bus boards are the Region of Interests for our model. The process involved in building the model for bus board detection module using transfer learning is shown in Fig. 5.

In order to ensure that the trained custom models have better detection accuracy, transfer learning from a pretrained YOLO v3 model was involved in the training. Transfer learning helps us to construct precise models in a timesaving way. Using transfer learning instead of starting afresh, we start from patterns that have been learned while addressing a different problem. Pre-trained model are used to initiate the transfer learning process. A pre-trained model is the one that is used to address a problem which is similar to a problem that we want to work out. A lot of computational costs are involved in training such models therefore models from well established literatures are used (e.g. MobileNet, VGG, Inception, YOLO) as shown in Fig. 6.

Fig. 4 Sample training images

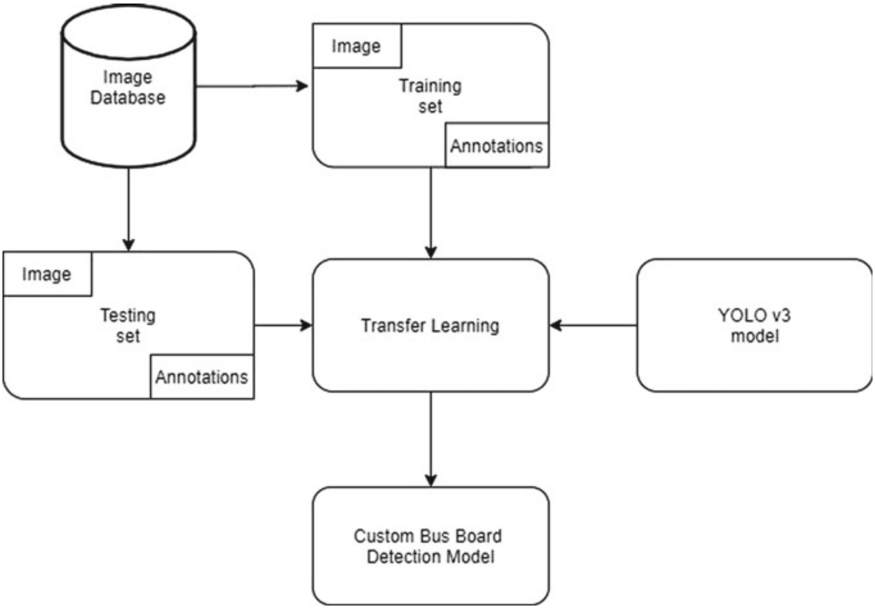


Fig. 5 System Architecture of proposed bus board detection module

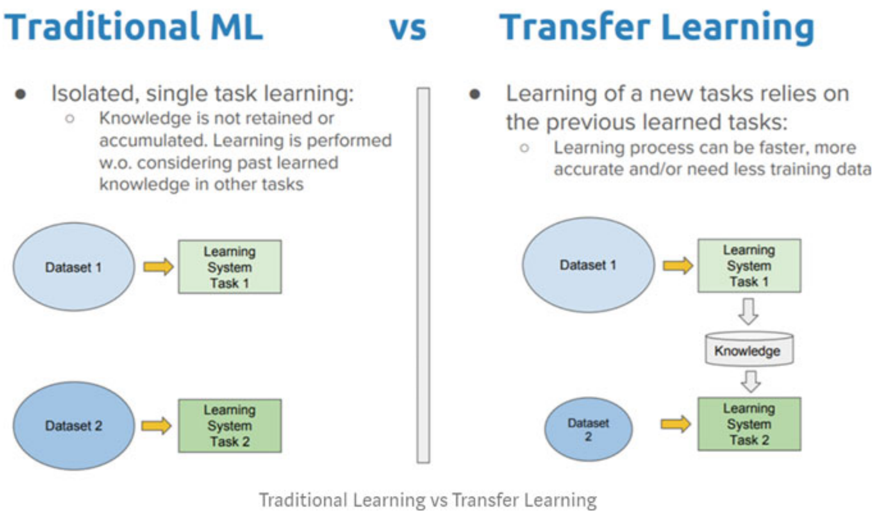


Fig. 6 Traditional ML versus transfer learning (*Source* https://towardsdatascience.com)

You Only Look Once (YOLO) is a fully convolutional network model and its outputs are generated by applying a 1×1 kernel on a feature map. In YOLO v3, the detection is done by applying 1×1 detection kernels on feature maps of three different sizes at three different places in the model network. These places are known as strides. Each of these strides are used for processing large, medium and small images respectively which is shown in Fig. 7.

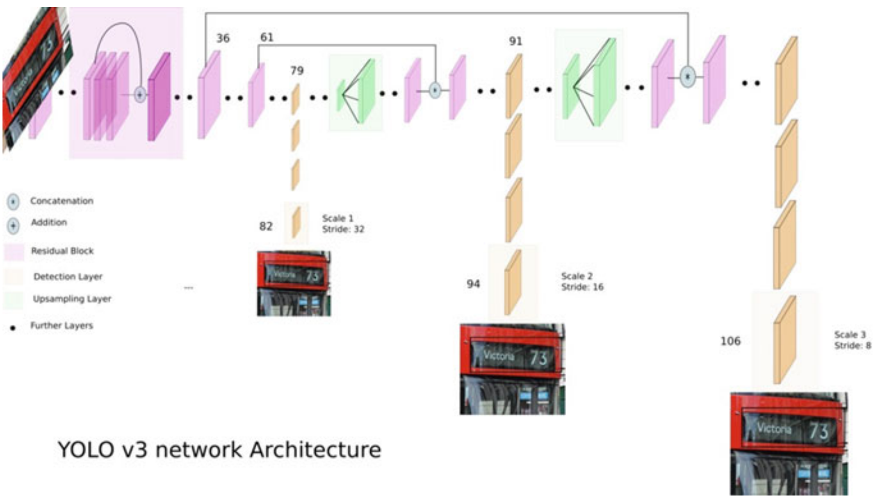


Fig. 7 YOLO v3 network architecture (*Source* https://towardsdatascience.com)

Transfer learning with annotated images was completed along with a pretrained YOLOv3 model. The obtained customized model is saved for later usage in bus recognition and bus board identification. After training is done, the model is evaluated for accuracy. They are identified based on the decrease in the validation loss. In most cases, the lower or less is the loss, the more accurate the model will be in detecting required objects. However, some models may experience overfitting and have lower losses. Therefore, to ensure that the best model is picked for our custom detection, we evaluate with mAP (mean Average Precision) of all the trained models saved in the detection folder. The detection accuracy of the model is improved with better mAP.

4.1.1 Bus Object Detection

The detection mechanism in a deep learning module mainly has two phases. In the first phase, an image is taken as an input and a number of blocks or boxes are formed around the possible entities with statistical features to label objects. Then this output is predicted to detect the object with its class name.

The feature to identify the bus object comes from the pretrained YOLO v3 model which has a Convolutional Neural Network as underlying layer and there is no requirement to train the bus images separately for its identification process. The initial phase consists of a number of convolution layers and pooling layers through which the image pixel values move in an array format. Activation functions are

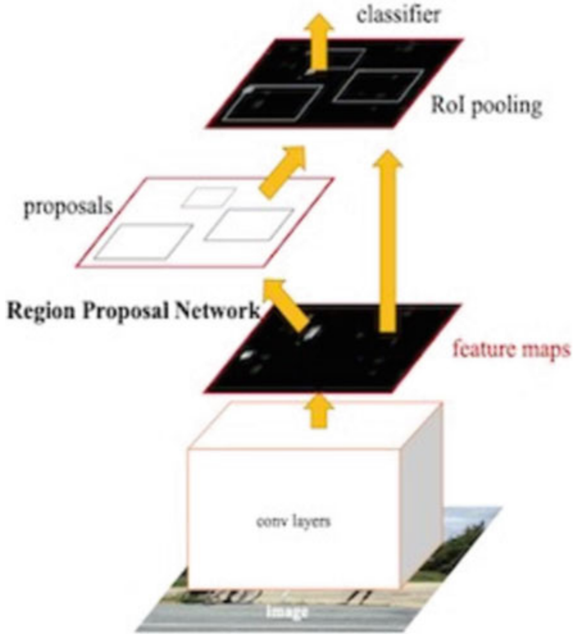


Fig. 8 Detecting ROI from random image (*Source* https://heartbeat.friz.ai)

Fig. 9 **a** Case 1—No buses recognized. **b** Case 2—Two buses recognized

present in the middle and through all of these, feature learning takes place. The next phase has flattened and fully connected layers with weights and before the final output classification is given soft maxing is done as shown in Fig. 8. The system moves to next processing only when the bus object is present in the video frame. Results of bus detection module is shown in Fig. 9.

4.1.2 Bus Board Segmentation

Our customized model is used for detection of bus boards from the video frames. For every frame, the detection algorithm is performed. The algorithm initially recognizes

Fig. 10 Sample output of detected ROI (bus name board)

the bus. Then in the image, segmentation as 13 $\times$ 13 grids is done and each cell has a maximum of five bounding boxes. Prediction for each of the bounding box areas is made and scores are provided. If it is identified to be below the threshold value from the trained model, they are neglected. For the areas with overlapping bounding box boundaries, features are collected together and prediction scores are given. Along with them, confidence score is also present which gives distinction between that class (ROI) and rest of the background image. The bus board (Region Of Interest) is finally detected if it is present in the image. If there are multiple bus boards in the frame, the same technique is applied and all of them are extracted and they are saved as separate grayscale images which are used for the next operation i.e., Optical Character Recognition (OCR). It is necessary that the detected bus board image must not be blurry or dark as it makes it difficult for OCR operation. The results of segmented bus board is shown in Fig. 10.

4.2 Route Recognition

For obtaining text from an image. Tesseract OCR is used in our system. The latest recognition engine of tesseract is based upon Long Short Term Memory (LSTM). For the system pytesseract module is chosen as it acts as a replacement for the command line tesseract with specified configuration arguments. As all of the other modules are python based, we also make use of Tesseract OCR in python as pytesseract.

Tesseract makes a two step approach for text detection. It makes use of adaptive recognition where in the initial step it identifies the pattern in letters, pixels, words and sentences, hence recognizing the characters. If it is not sure or successful in finding some characters, in the next step it tries to fill in with the character or word that matches the word or sentence context. In character recognition step the outlines are arranged into blobs by the nesting process. These blobs are then structured into lines of text and each of these text lines are broken into possible words and further into characters as shown in Fig. 11. Multiple passes are involved before finally predicting the character, words and the sentences. The process involves a number of iterations before final text is produced. In each iteration, the accuracy of recognition improves and finally finely converted text data is obtained.

In our system all the information present in the bus board image is obtained by operating the tesseract tool on the image. This obtained text data is stored as string

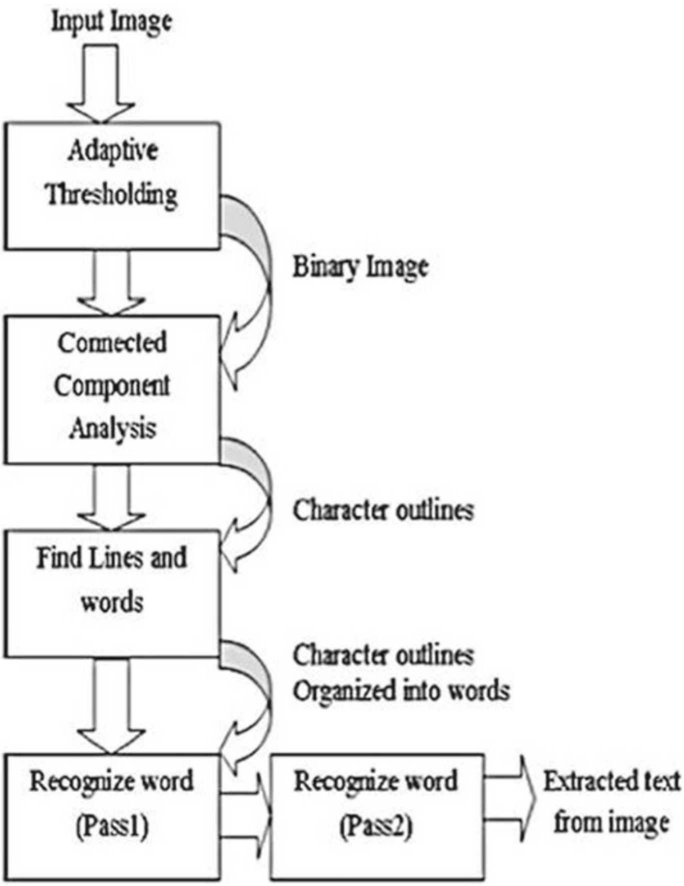


Fig. 11 Working of Tesseract OCR (*Source* https://blog.cedric.ws)

type data for further processing. For obtaining text with good accuracy, quality images are required from the previous module.

4.3 Text to Speech Conversion for Route Identification

Regular expressions are defined to obtain the numeric or alphanumeric bus number from the text data stored in a text data format. A regular expression is a special text string for describing a search pattern. Strings of text are compared to the pattern in order to identify the string that matches the logical pattern defined by regular expression. On the basis of these comparisons the regular expressions can be used to identify strings of text that meet specific requirements or to validate that strings meet a required pattern. It is very important to define regular expressions properly because even if a single letter or number is missing, the route number becomes ambiguous. The process involved in extracting the bus route and converting it in to speech output is shown in Fig. 12.

The constraint laid for our regular expression is to work in a way such that it only extracts maximum 3 characters that are either numeric or alphanumeric and are continuous characters present one after another. The constraint is based in such a way because the bus numbers in the modern transport system are in numeric or in alphanumeric format and their characters length are usually two or three characters long. This constraint can be adjusted based on the locality requirements.

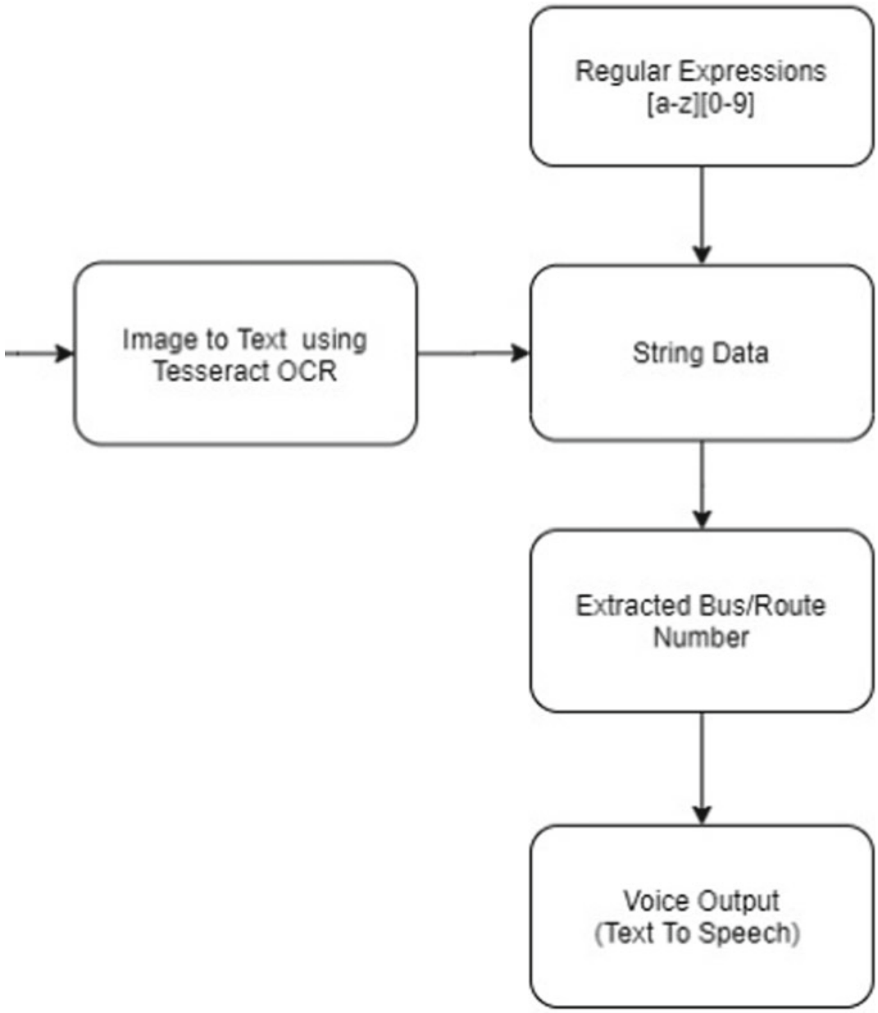


Fig. 12 Process involved in bus route identification

Pseudo Code for Bus Number Identification:

```
i- input image
t- String variable
ROI- Region of Interest
Img- Bus board image
RE- Regular Expression
Num- Bus Number
    1.  Initiate Function Identify_busnum()
    2.  i ← Capture()
    3.  Img ← ROI(i)
    4.  If data not in Img:
            a.  Suggest Capture()
    5.  Else:

            a.  t ← OCR(Img)
            b.  Num ← RE(t)
            c.  Voice (Num)
            d.  Query_db()
    6.  End!
```

An additional module, a database can be added to the system so that with the extracted bus number, all other information regarding the bus can be provided to the user. The bus number serves as the primary key with which the database is queried to provide details like destination location, stops present in the middle, ticket cost and other necessary information.

4.4 An IOT Based Railway Platform Detection

This subsystem involves detection of railway platform when train stops at the stations. An ultrasonic sensor is mounted to the headwear like cap/hat at a certain angle facing the ground. Initially the user's height data is loaded in pi memory. This height along with a small variable length is added together and the value is stored as a prefixed threshold value. The variable length is the difference in height from platform to the train door. An ultrasonic sensor measures the distance based on the working given in Fig. 13.

The transmitter sends out waves which travels at the speed of sound waves (0.034 cm/μs) and the receiver obtains them when they are reflected back by an object or surface that is present in their path. The time difference (t) is noted and along with the known speed of the sound, distance value can be calculated (v*t/2) in the pi. The algorithm for the platform detection mainly identifies whether this measured distance exceeds the prefixed threshold value. If so, the user is notified quickly through speakers to get down on the other side of the train. Ultrasonic sensor is very much responsive therefore this operation takes place quickly. It is a light weight and simple process to assist the blind for safe exit from trains. The same methodology can be extended in future also for other transports. Some ambiguity may arise if there are obstacles present between the sensor and the ground/surface.

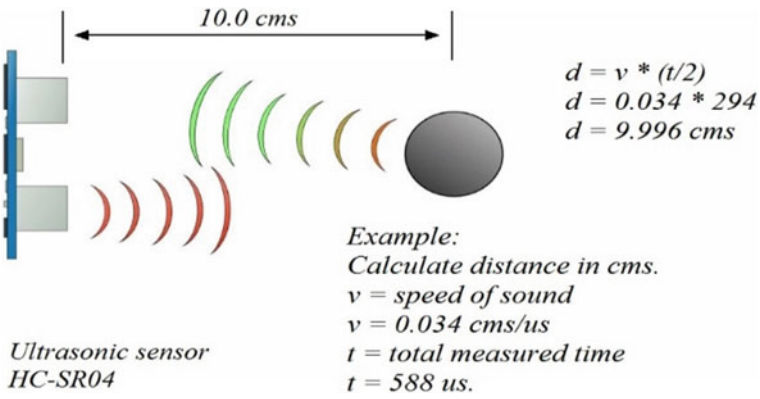


Fig. 13 Working of an ultrasonic sensor (*Source* https://howtomechatronics.com)

This can be overcome by adding multiple ultrasonic sensors to the system as shown in Fig. 14.

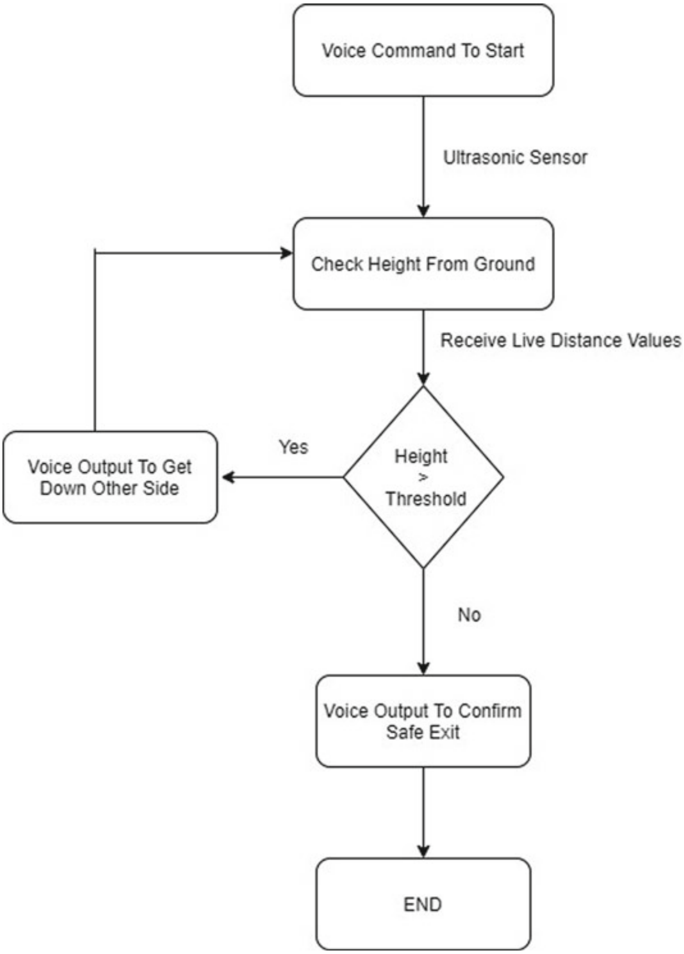


Fig. 14 Flow diagram for platform detection module

Pseudo Code:

```
t- threshold distance
m- measured distance

1.  Initiate function ultrasonic_measure(t)
2.  While(true):
            Find m
        a.  If m>t:
                i.   Voice("danger")
                ii.  Continue
        b.  Else:
                i.  Exit Loop
3.  End!
```

4.5 Assistive Device for Visual Mobility and Safe Travel

A single integrated system is developed and the whole system is attached to a cap to assist the visually impaired. User can choose the required operation by commanding through voice or using a button. Raspberry Pi serves the processing unit for all the operations and it is connected to an external portable power supply.

For bus identification, a High Definition wide angle camera is used and the real time feed starts once bus number identification operation is chosen to obtain the bus number and all other details.

For railway platform detection, an Ultrasonic Sensor is present at a certain angle facing ground every time. It is also connected to the raspberry pi and it gives the frequent depth measures from ground to the pi in which the depth identification algorithm is executed. Speaker/headphones are present to give output notifications. Microphone is used to obtain the voice input from the user. If the voice input is found to be difficult for the user or if there is lot of noise, button inputs can be added to the system. GPS can be added as an extended feature which can inform the user where they are currently located and to know the remaining time it would take to reach the destination as shown in Fig. 15.

```
timeout = pulse_end + maxTime
while GPIO.input(ECHO) == 1 and pulse_end < timeout:
    pulse_end = time.time()

pulse_duration = pulse_end - pulse_start
distance = pulse_duration * 17000
distance = round(distance, 2)

print(distance)
if distance >=120:
    espeak.synth("Danger! Danger! ")

    time.sleep(5)
    print("Danger!! Exit Other Side")
```

Shell

```
28.28
28.02
27.49
227.31
Danger!! Exit Other Side
26.27
```

Fig. 15 Results of IOT based railway platform detection

5 Implementation and Results

More than 400 images were chosen as a dataset to create a customized model to detect the bus board. The initial accuracy was computed from the ability of the model to find whether buses were correctly recognized in the images (i.e. for bus recognition). Thirty traffic images with vehicles were taken and the model was made to identify whether buses were present in the image or not. Out of thirty predictions twenty seven right predictions were made by the system. Hence the accuracy was found to be 0.9 out 1 i.e. 90%.

Fifty random bus images were chosen in order to test the accuracy of the bus board detection model. Although there were cropped edges, the region of interest was properly extracted. The accuracy is identified by calculating the ratio of number of properly extracted images to the total number of images chosen. The accuracy was found to be 0.74 out of 1 i.e. 74% with the native algorithm. The accuracy improved by 6% for the same dataset by enhancing the algorithm which includes an increase in the number of anchor boxes in the cells of the image. The results of segmented bus boards are shown in Fig. 16.

Compared to accuracy of models proposed using HOG, SVM, and CNN, our system has been identified to be more accurate. It has also been noted that the proposed solution is quick as it makes use of the power YOLO object detection model. The comparison is based on accuracy of bus board detection as shown in Table 1 and Fig. 17.

Fig. 16 Detected bus board images

Table 1 Comparative performance analysis

Method	Total no. of images used for identifying accuracy (a)	No. of properly processed images (b)	Accuracy b/a(%)
HOG and SVM	50	36	72
Yolo	50	37	74
Enhanced Yolo	50	40	80

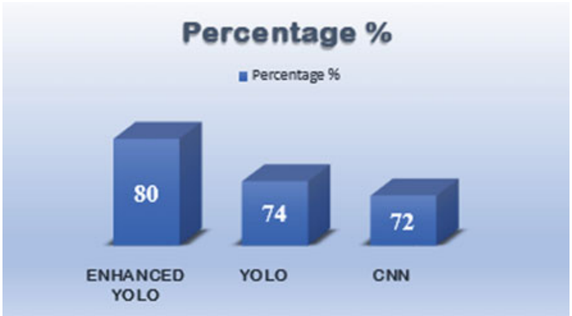


Fig. 17 Comparative performance analysis of proposed system

6 Conclusion and Future Work

In this paper, we propose a system that helps visually impaired mobility in bus stations in real time. We hope with this kind of system, their lifestyle can be improved and their stress from fear of meeting with any kind of accident while travelling can be reduced. The system proposed can be integrated with existing devices or they can be implemented as a separate one. GPS and location tracking can be further added so that the user can know where they are exactly, distance remaining and time that will be taken to reach the destination. The system can be connected with the smart phone for booking cabs or other transportations and inform the user when the vehicle arrives by identifying the vehicle register number or if any QR code/Magnetic Tags are present in the vehicle.

References

1. Wongta, P., Kobchaisawat, T., Chalidabhongse, T.H.: An automatic bus route number recognition. In: Proceedings of the IEEE International Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, pp. 1–6 (2016)
2. Guida, C., Comanducci, D., Colombo, C.: Automatic bus line number localization and recognition on mobile phones—a computer vision aid for the visually impaired. In: Maino, G., Foresti, G.L. (eds.) Image Analysis and Processing—ICIAP 2011, pp. 323–332. Springer, Berlin, Heidelberg (2011)
3. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
4. Pan, H., Yi, C., Tian, Y.: A primary travelling assistant system of bus detection and recognition for visually impaired people. In: IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6 (2013)
5. Tsai, C.M., Yeh, Z.M.: Text detection in bus panel for visually impaired people "seeing" bus route number. In: 2013 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 03, pp. 1234–1239 (2013)
6. Bouhamed, S.A., Eleuch, J.F., Kallel, I.K., Masmoudi, D.S.: New electronic cane for visually impaired people for obstacle detection and recognition. In: Proceedings of the IEEE International Conference on Vehicular Electronics and Safety, pp. 416–420 (2012)

7. Zahir, E., Hossain, M. S., Iqbal, M.W., Jalil, I., Kabir, S.M.: Implementing and testing an ultrasonic sensor based mobility aid for a visually impaired person. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (2017)
8. Ani, R., Maria, E., Jameema Joyce, J.: Smart specs: voice assisted text reading system for visually impaired persons using TTS method. In: IEEE International Conference on Innovations in Green Energy and Healthcare Technologies (2017)
9. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: The 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3538–3545 (2012)
10. Gómez, L., Karatzas, D.: Multi-script text extraction from natural scenes. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 467–471 (2013)
11. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
12. Kobchaisawat, T., Chalidabhongse, T.H.: Thai text localization in natural scene images using convolutional neural network. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, pp. 1–7 (2014)

Dr. R. Priyatharshini received her BE degree from University of Madras in 2001 and ME degree from Anna University in 2008. She has completed her Ph.D. degree from Anna University in 2017. She has 18 years of teaching experience and currently working as an associate professor in Easwari Engineering College, Chennai. She has published 26 research papers in International Conferences and Journals. Her Area of interest towards data science, machine learning, deep learning and medical image analysis. She is a member of Institution of Engineering and Technology. She received Best Project Award by TNSCST in 2017 for guiding the project titled “Demand based Blue Collared Job Allocation using Multiple Priorities and Heuristics” and also received Outstanding Researcher Award from SBYTE Educational & Research Foundation in 2020.

Dr. R. Senthil Kumar completed Bachelor of Engineering with Honors (B.E. Hons.) in the year 1984, from PSG College of Technology, Coimbatore, affiliated to the Madras University. He was awarded Master of Technology (M.Tech.) with Distinction from Indian Institute of Technology, Delhi (IIT-Delhi) in the year 1986. He was awarded Ph.D. in the year 2010 by Anna University, Chennai for an inter-disciplinary thesis on database, XML and data migration for industry applications. He has more than 35 years of experience of which more than 15 years in Industry and 18+ years in Academic Institutions as a Head of the Department/Principal. He worked with M/s. Hindustan Motors Ltd. (MannesMann Demag BauMaschinen Division), M/s. Ashok Leyland Ltd. (Product Development Division), M/s. HCL Ltd. (CAD/CAM Division), M/s UMS Services at Coimbatore and as a Director at M/s. IIDT (P) Ltd., a software development company. He has provided leadership to the different Institutions in Strategic Plan, R&D, Industry-connect, Quality Assurance in Education. He is Program Evaluator for NBA. His domains of interest are OBE, database systems, CAD, CAM, ERP, PLM, data interoperability, Quality in Engineering Education, pedagogy. He was instrumental in different Institutions to get more than Rs. 15 crores as project funds and published 28 technical papers. Dr. R. Senthil Kumar is presently working as Professor and Principal at Easwari Engineering College, Chennai.

M. Sanjay Sivakumar completed his B.Tech. in Information Technology. His area of interest includes Computer Vision, Deep Learning and Machine Learning. He is currently working as software engineer trainee under retail domain in Tata Consultancy Services Ltd. He was a member of Rotaract Club of SRM Easwari Engineering College and has done many services as a Public Relations Officer of the club.

A. Mathumathi completed her B.Tech. Information Technology from Easwari Engineering College. Her area of interest is towards Machine Learning and Deep Learning. She is currently working as Graduate Engineer Trainer in a financial service company TVS Credit Services Limited. She was a member of Rotaract Club of SRM Easwari Club of Easwari Engineering College and organised many events for social cause.

Nancy Sharon Johnson completed B.Tech. degree in Information technology from Easwari Engineering College. Her area of interest is towards Machine Learning and Deep Learning. She is currently working as a Graduate Engineer Trainee in a L&T Infotech. She was a member of Youth Red Cross Club. She acted as a volunteer in many social service activities.

Soft Computing Techniques for Physical Layer Security of IoT Devices

C. Ismayil Siyad and S. Tamilselvan

Abstract Internet of Things (IoT) is an archetype of the Internet where the devices are connecting each other utilizing the Internet. Due to the enormous attention from industries and academia IoT is considering one of the technologies that revolutionize the future. IoT is considering as one of the potential enabler technologies for beyond fifth-generation wireless technologies. The physical layer aspects of IoT is one of the rapidly developing areas in IoT research. The developments in the arena of soft computing techniques such as fuzzy logic, genetic algorithms, machine learning, and deep learning algorithms have an unequivocal role in this buildout. Data-driven techniques revolutionized various aspects of physical layer techniques such as the generation of adaptive waveforms, energy harvesting, the energy efficiency of the network, spectrum sensing, multiple access techniques, cooperative communication, power allocation, and carrier allocation, etc. Physical Layer Security (PLS), where communication security is achieved by the techniques used in the physical layer is a major application area of soft computing techniques. Several PLS techniques for IoT based on soft computing techniques are proposed by various researchers. Soft computing techniques transformed many of this hardware define techniques to software-defined. In the literature, many researchers are recently reporting many advancements in this domain. One of the major advantages of this type of data-driven ad knowledge-driven technique is its inherent ability to adapt and cognitive capacity to behave differently with the time-varying characteristics of the physical medium. Another advantage of soft computing technique in the physical layer is its propensity to solve nonlinear problems which are difficult to solve with mathematical algorithmic models and its ability to approximate complex dynamic systems according to the renowned universal approximation theorem. Also, for some solvable multivariate optimization problems soft computing techniques give low complexity solutions by training the same algorithm generated data. Thus, creating a low complexity representation of nonlinear models. In this regard, this chapter presents a comprehensive overview of

C. Ismayil Siyad (✉) · S. Tamilselvan
Pondicherry Engineering College, Puducherry, India
e-mail: ismayilsiyad@pec.edu

S. Tamilselvan
e-mail: tamilselvan@pec.edu

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_2

the state-of-the-art approaches towards the application of soft computing algorithms to physical layer security techniques for IoT network. Qualitative and quantitative insight into soft computing techniques for the IoT physical layer security is included in this chapter.

Keywords IoT · PHY · Physical layer security · Soft computing · Machine learning · Deep learning · Physical layer security · Fuzzy logic · Genetic algorithm

1 Introduction

IoT is considered to be the potential facilitating technology for 5G and beyond wireless networks. IoT transforming the Internet from being the network of 'people' to network of 'Things'. It provides the capability to sense and communicate with the 'Things'. So that objects can have to internetwork with objects. This type of Information Communication Technology (ICT) is revolutionizing the way people and objects are interacting. Many protocols are in use for IoT communications which are 6LoWPAN, Zigbee, Bluetooth Low Energy (BLE), Near Field Communication (NFC), and SigFox, etc. Numerous physical layer techniques are employed in all these protocols. Further, for long-range communications, IoT uses Cellular and LoRa technologies. In the development of communication technology for IoT, the major concerns should be light weightiness, adaptability, and energy efficiency of the proposals. Being an unequivocal ally of human life, the major challenge facing by the IoT is the security of the data, which is handling by the network. Apart from the personal application IoT have significant effects on different areas such as agriculture, production, smart-grids, home, and industrial automation, health care, and automobile, etc.

The maneuvering data through these systems are vital, which makes securing this data is so crucial. But the security of IoT is in its infant state. One of the key challenges to secure IoT communication is its inept hardware doesn't allow advanced cryptographic algorithms to covert the message that being transmitted from the IoT devices. Also, the connected 'Objects' and traffic generated in the network are massive in number. Providing cryptographic encryption at the higher layer is for each of these devices is not at all feasible. Also, the lack of standards and best practices make the issue far-reaching. Thus, the security breaches in IoT is exorbitant. Consider the case of health care equipment like a cardiac pacemaker, A Denial of Service (DoS) attack on it can cost a life. In all the above-mentioned applications the breach of security can make serious defacements to its cause. Considering the inappropriateness of the chunky cryptographic algorithms on IoT motes, the feasible security solution for IoT is the PLS.

The PLS is a recently evolved security paradigm. The foundations of PLS gleaned from the principles of information theory. Here the security is considered in the physical layer of the system. The principal cause for the ubiquitous acknowledgment of PLS as the security solution of the future is due to its inherent ability to cope up with the computationally well-resourced eavesdropper. This is since the security in the PLS does not cling to the computational ineffectuality of the eavesdropper. The cryptosystems used for traditional data security is based on the mathematical exercise. Instead, the data security is ensured by the physical characteristics of the channel between the legitimate users and that of the intruder. The fundamentals of PLS arise from the earlier works of Claude E Shannon. He proposed the perception of perfect secrecy by considering the existence of a shared secret key which is unknown to the eavesdropper, subjected to the fact that the entropy of the shared secret key is larger than the entropy of the message transmitted. The PLS leveraged the innate characteristics of the wireless communication systems. For example, it capitalizes on the random nature of the channel variables such as channel gain, channel phase, and noise, etc. A multitude of techniques is proposed in the literature for the practical implementation of PLS. It can be broadly classified into Coding, Multi-antenna technique, artificial noise injection, cooperative communication, and secure key generation, etc.

The conventional cryptosystems are part of many existing IoT communications protocols stacks. The BLE adopts the Elliptic Curve Diffie-Hellman (EC DH) for secret key sharing and substantiation of the authenticity of users. LoRaWAN employs a network key and application session key for the same purpose. The Zigbee standard exercises network key and link key for the authentication of legitimate users. Due to the aforementioned hindrances of conventional cryptosystems, their application is a resource-constrained IoT system that is further an arduous task. This situation awakens the need for a new security paradigm, where nuance lies in the fact that the computational incompetence and outfitted eavesdropper's computational completeness will not matter. This gap can be perfectly filled by the usance of PLS techniques in IoT. An ample amount of PLS techniques are proposed in the literature. The state-of-the-art techniques can be categorized in terms of encryption-based, artificial noise based, compressive sensing based, etc. In a compressive sensing-based approach, the sparsity of the message signals [1, 2] is exploited for the linear transformation of the message data with the sensing matrix or measurement matrix. Here, the seed for generating the sensing matrix is based on the quantization of the physical characteristics of the channel. Another method in literature is the usage of cooperative communication techniques for the PLS. This technique is more viable for IoT architecture. Due to the massive number of smart objects forming the IoT networks, the trusted nodes inject artificial interference to avoid the reception at the wiretapper [3, 4].

Another approach for cooperative communication-based PLS in IoT is the volition of the relay nodes. The fundamental approach of PLS in every network is the injection of artificial noise. The injected artificial noise will deteriorate the reception at the eavesdropper. At the same time, the reception at the licit user is less affected. The basic detriment of this approach when employing in IoT network is that the requirement of multiple antennas in the nodes is not viable for the resource-limited IoT nodes [5, 6]. To confront this issue cooperative artificial noise injection where numerous nodes take part in the noise injection in the desired direction. An alternative approach in PLS for IoT is physical layer encryption based on the key generation [7, 8]. Here the key is generated based on the common source of randomness between legitimate users, usually the intrinsic channel conditions through probing, quantization, information reconciliation, and privacy amplification. The data is encrypted using the generated key. The received message can be retrieved by the legitimate user since the key is exactly reproducible at the legitimate user site but it is unachievable for the eavesdropper. Another methodology for PLS in IoT is the design of PLS codes most favored contender in this category is Low Density Parity Check (LDPC) codes and its variants. Many other practical codes such as polar codes and lattice codes are also proposed in the literature. The major challenge in deploying PLS codes in IoT networks is its computational complexity that is surplus for the energy and cost constraints of IoT. Many proposals are putting these IoT constraints first, such as constellation rotation techniques, lightweight codes like raptor codes [9], Luby Transform (LT) codes [10], and Reed Solomon (RS) codes [11], etc.

The sophisticated pragmatic problems cannot easily deal with mathematical modeling or hard computing due to its ambivalent, time-varying, and nonlinear nature. To address the aforementioned problems, methods having andromorphic intellectual which are more resilient to dynamic nature. The collective of these artificial intelligence techniques can be categorized as knowledge-driven techniques such as Fuzzy Systems (FS) and data-driven techniques such as Machine Learning (ML), Genetic Algorithm (GA), and Deep Learning (DL) as shown in Fig. 1. These methodologies are applied in their intrinsic form for solving a specific type of problem, whereas the fusion of techniques works well for some other kind of problems. The properties of soft computing techniques such as cognition, learning, understanding, and adaptability make these techniques vigorously effective and articulate for numerous arenas of real-world problems. In wireless communications, soft computing techniques have a consequential contribution to revolutionizing the diverse segments of wireless communication techniques. ML and DL techniques have

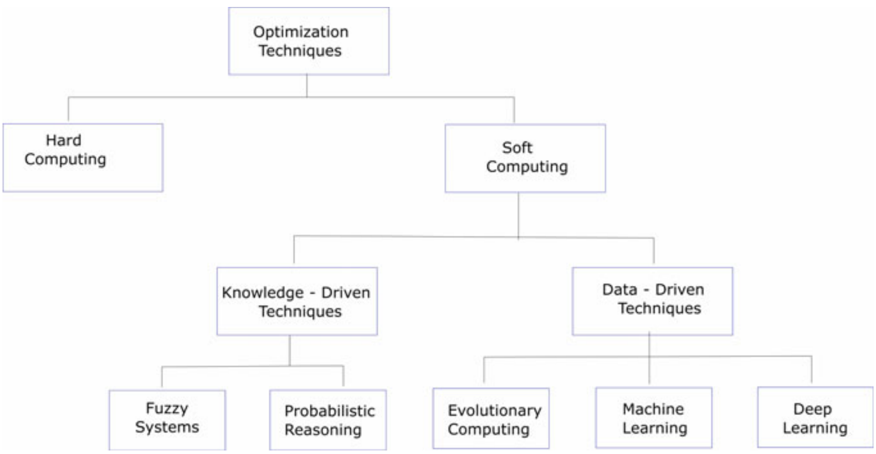


Fig. 1 Categories of problem solving and optimization techniques

prolific applications in different sections of wireless communications techniques and IoT communication techniques such as, channel estimation [12, 13], Channel State Information (CSI) feedback [14], PLS [15], DoA estimation [16], location estimation [17], Non-Orthogonal Multiple Access (NOMA) [18], and Channel equalization [19], etc. Also, fuzzy systems have significant contributions in various aspects of wireless communications such as channel estimation, channel equalization, decoding, channel access, power control, Power allocation, and resource allocation and PLS. The usage of soft computing techniques for PLS in IoT networks is the potential area of interest in this chapter. The potential and prospects of soft computing techniques, when it's been using for the PLS technique in the IoT network are comprehensively summarized in this chapter emphasizing the following aspects.

1. Applications ML algorithms for PLS in IoT.
2. Applications of DL algorithms for PLS in IoT.
3. Applications of FS for PLS in IoT.
4. Applications of other soft computing techniques for PLS in IoT.

2 Machine Learning Algorithms for PLS in IoT

The learning process of ML algorithms can be enlisted under two pathways, classification methods and clustering methods. The examples for classification methods are K-nearest neighbor (KNN), Support Vector Machine (SVM), Decision Trees (DT), Linear Classifiers (LC), and Random Forest (RF), etc. The KNN is a supervised learning Scheme in which the unlabeled data is classified based on the label of the nearest neighbors. Here, the proximity of the data is considered as the similarity of the data points. Another supervised learning algorithm SVM works based on the selection of the best hyperplanes as a decision based on the selection of best possible hyperplanes as decision boundaries that separates different classes and the decision is made based on the boundaries within which data point is entrapped in. Even though the major area of application of decision trees is in supervised learning it can also be used for unsupervised learning. The decision trees disintegrate the data points into subsets and an assorted tree is formed. The classification decision is based on the tree structure.

2.1 PLS Schemes Using K-Nearest Neighbor Algorithm

Li et al. [20] proposes a KNN based eavesdropping detection system that exposes the nodes with unusual activities as depicted in Fig. 2. The K-value considered is less than the number of intruder nodes, also the study further expands to the situation when K value equals and greater than the same. The trend followed by the system is that as K value increases the performance is better. Distributed Denial of Attack

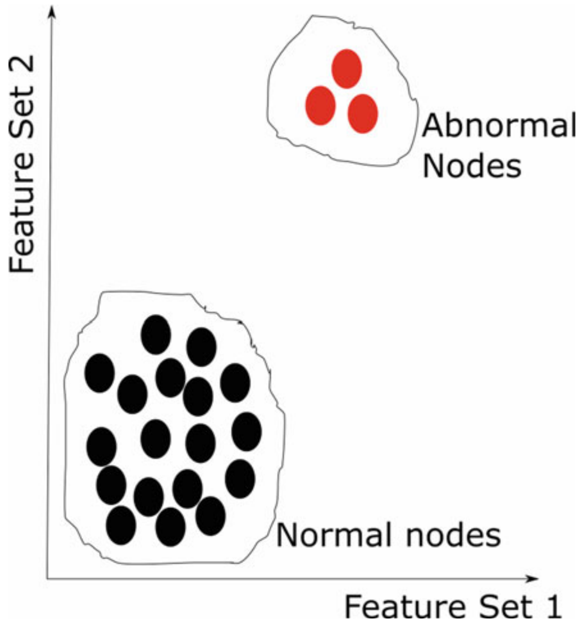


Fig. 2 KNN based intruder detection algorithm

(DDoS) is considered as the attack model for the analysis, which is the most common and very severe in IoT networks.

The physical layer authentication based on KNN and SVM is studied by Senigaglesi et al. [21]. The proposed work considers the presence of passive eavesdropper considering the time-varying fading channel. The study highlights the resilience of the clustering-based and neural network-based model over the One Class Classification (OCC) equivalents. The work explores the effect of the size of the data set and they conclude that larger data sets will not have the merits considering the PLS perspective using ML. The performance metrics chosen to assess the prominence of the proposal are the probability of false alarm and the probability of missed detection. The trade-off between these two parameters is optimized using these data-driven techniques. These metrics are assessed by exploiting the confusion matrix as shown in Table 1.

The probability of missed detection and the probability of False Alarm (FA) can be determined as

Table 1 Confusion matrix for the analysis of intrusion detection scheme

	Predicted class		
Real class	+	+	–
		TP	FN
	–	FP	TN

$$P(missed\,detection) = \frac{FP}{FP + TN} \tag{1}$$

$$P(false\,alarm) = \frac{FN}{FP + FN} \tag{2}$$

The performance metrics for the eavesdropping system proposed in [22] are the detection rate and false alarm rate, which are 98.5% and 4.63% respectively. Fei Pan et al. [8] studied the application of machine learning techniques such as KNN, SVM, DT, and Ensemble Learning (EL) for PLS in wireless industrial cyber-physical systems as shown in Fig. 3. Industrial IoT (IIoT) is a revolutionary technology that has manifold effects in the production and supply chain. These connected industrial objects need to be secured without increasing the network overhead. PLS is an apt technique that can achieve this. This work proposes a physical layer encryption scheme based on the estimated CSI at the wireless nodes. Apart from typical KNN and SVM based classifiers, this work also deploys the DT method and EL technique

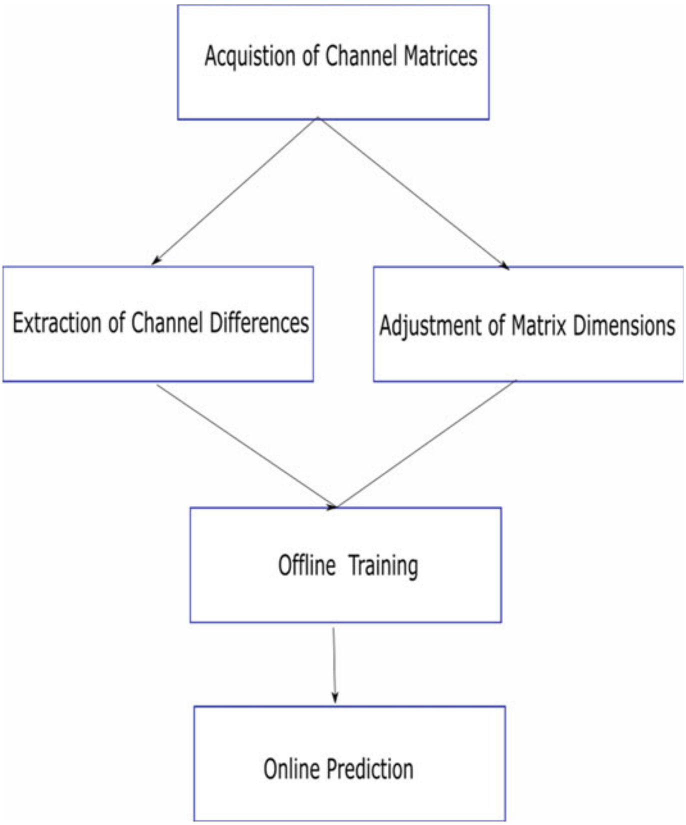


Fig. 3 Framework for ML based Physical Layer Security

for PLS. The DT tabulates the CSI values to authorized or unauthorized node data using simple tree configuration and pruning with logical reasoning using the "if-then" questions. The KNN distance metric is chosen is the cosine distance.

$$d(cosine) = \frac{h_t \cdot h_{train}}{||h_t||^2 \cdot ||h_{train}||^2} \tag{3}$$

The SVM classifier used for this work is a multi-classifier SVM algorithm namely Coarse Gaussian SVM (CG-SVM) having the Gaussian kernel function.

$$k(h_i, h_j) = \frac{||h_i - h_j||^2}{2B^2} \tag{4}$$

where B is the nonzero bandwidth of the kernel.

The ensemble learning technique deployed in this work is bagged tree meth-od. In which, the batch of base simple trees with the individual base decision and the final decision is chosen based on the voting process as shown in Fig. 4. The simulation is carried out based on the National Institute Standards and Technology (NIST) automotive factory dataset. The performance is analyzed taking authentication accuracy and prediction time as the figure of merit. The BT method got the best accuracy rate comparing to the counterpart classifiers. The simulation results are validated using the empirical test results.

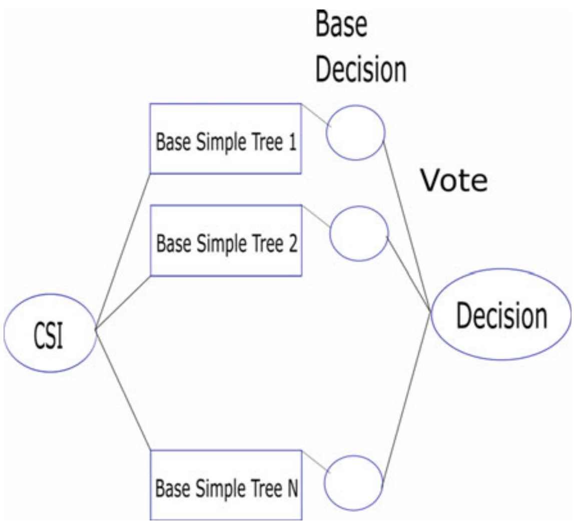


Fig. 4 Ensemble learning technique based on bagged tree algorithm

2.2 PLS Scheme Using Support Vector Machine

Hoang et al. [23] proffer a novel SVM based technique to detect the active eavesdropper in wireless communication. The network considered in this work consists of K number of legitimate users, one legal access point, and an active attacker. This network scenario is fully consistent with typical IoT network architecture. The work relies on the dataset generation framework. From the collected dataset unlabeled feature set is formed on the ideal features and practical features such as Mean, Ratio, and Sum. The labeled dataset for a supervised learning scheme is generated as Artificial Training Data (ATD). The Radial Basis Function (RBF) kernel is chosen

$$k(xs, x) = e^{-r||xs,x||^2} \tag{5}$$

where r is set to 0.5 and the accuracy of the different combinations of features are examined. The blend of Mean and Ratio feature combination found more robust. Four alternatives of SVM classifiers with 4 different kernels namely linear kernel, RBF kernel, polynomial kernel, and the sigmoid kernel is applied on a combination of feature sets. Various kernels applied to the fusion of Mean and Ratio features are picturized in Fig. 5. The dominance of the SVM method with the RBF kernel is established by the empirical results.

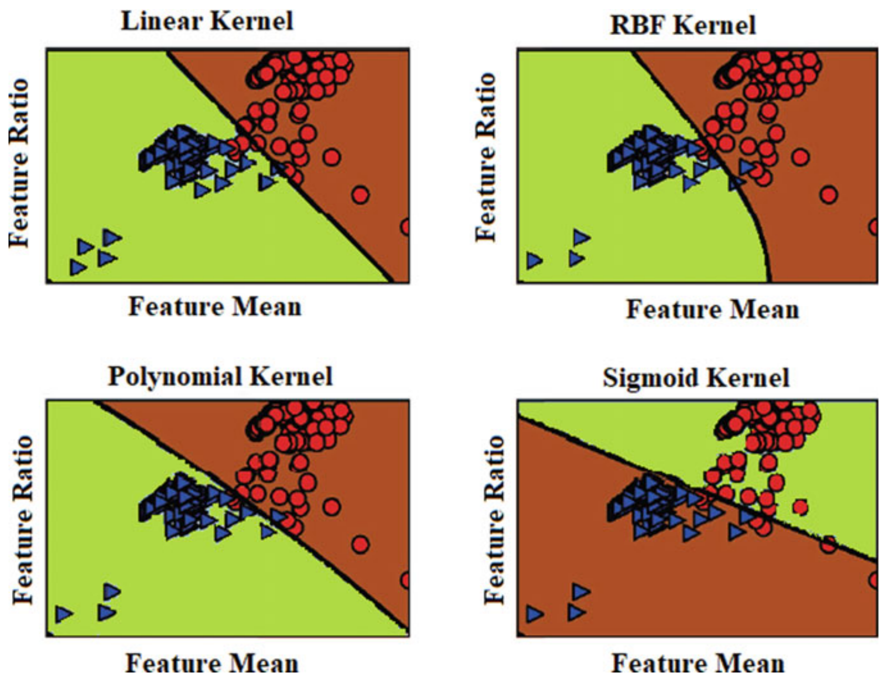


Fig. 5 SVM classifier for PLS with different kernels [22]

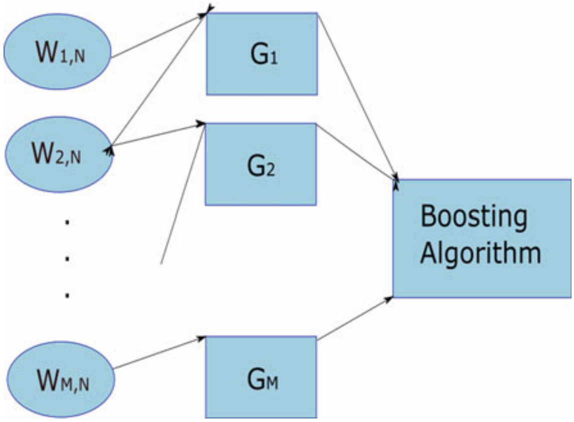


Fig. 6 AdaBoost algorithm for PLS

Chen et al. [24] proposed an Adaptive Boost (AdaBoost) based authentication methodology which can unmask the spoofing attack. The proposed scheme is befitting the resource limited IoT standard due to its lightness and low network overhead nature. The AdaBoost algorithm considers SVM as the weak classifier as shown in Fig. 6. At the training phase each misclassified data point is undergone weight boosting and subsequent phase of classification is carried out. These training steps are repeated until there is no misclassification. The AdaBoost algorithm based PLS is structured as shown in Fig. 7. The chosen classifier is the decision model which form the linear combination of trail classifiers forming a voting-based decision. A pragmatic MIMO model is formed and the practical aspects of authentication in physical layer is verified. The performance of the system is verified based on the decision rate metric for two different test scenarios named as T_A and T_B where the difference is in the overlaid area between frames of Eve and Alice with resonant frequency 3.5 GHz and bandwidth of 2 MHz. All the proposed systems imply the fact that ML based approach have significant improvement over the threshold-based detection approach in terms of detection rate.

Fang et al. [25] approached ML-based PLS using Kernel Machine (KM) blending a multitude of physical layer attributes without much apprehension on the underlying statistical properties as shown in Fig. 8. To lower the computational complicacy, the problem is formulated as a convex optimization problem. Considering the effects of time variability, such as the limited availability of time and other computational resources, the complexity of large search space, etc. The performance of the proposed system is analyzed based on the Mean Square Error (MSE) of authentication.

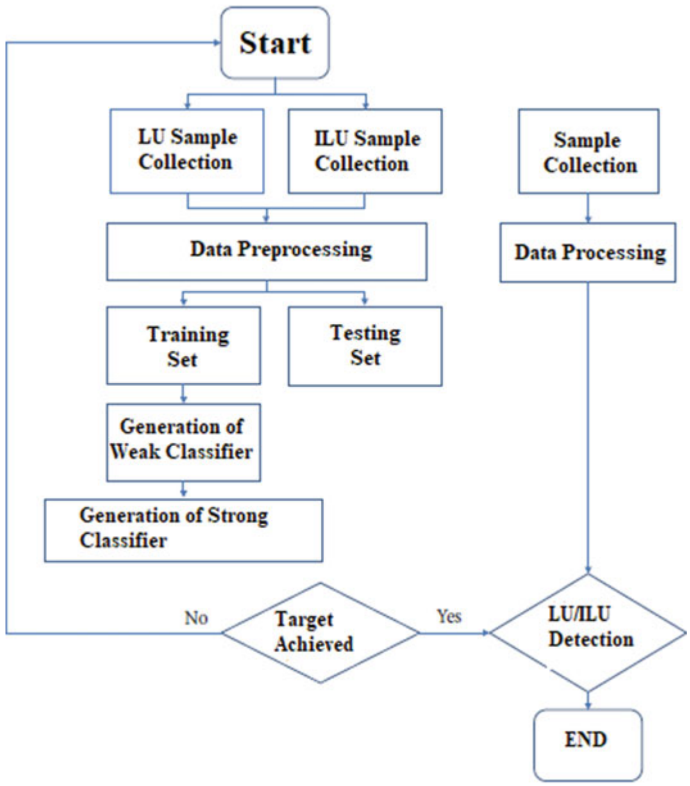
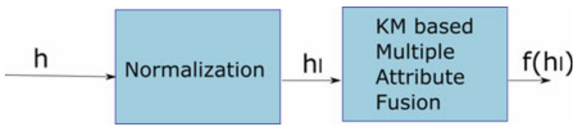


Fig. 7 PLS algorithm based on AdaBoost

Fig. 8 Kernel machine based physical layer attribute fusion



3 Deep Learning Algorithms for PLS in IoT

DL is an ML subset slanted towards the algorithmic model inspired by human brain functioning called Artificial Neural Network (ANN) with a greater number of hidden layers in between the input and output layers. The major objective of DL based algorithms is to exploit the bounteous real-life unlabeled or labeled data available. A productive cycle of more data, bigger models, and more computationally capable hardware provide better understanding and more accurate results for real-life problems with human-like expertise. ANNs, CNNs, Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) are the enduring DL models used in various areas of application. The most basic DL model is the Multi-Layer Perceptron (MLP) model and backpropagation model. The MLP model is a feed-forward neural network with weights and biases as the hyperparameters. Back Propagation (BP) is used to tune the hyperparameters using a gradient descent algorithm. The CNNs are a special type multi-layered feed-forward neural network which has wide applications in image processing and data analytics. The fundamental uniqueness of CNN is the presence of convolutional layers and pooling layers as the hidden layers. RNN and LSTM are memory-based approaches for DL. Here the sequential attributes of the data are exploited for the adaptation in the futuristic similar scenario. The mutual exclusiveness of the input-output relation is not holding in LSTM and RNN, which can add, delete, or change the data as necessary. The application of these DL models for the PLS in IoT based network is explained in subsequent sections.

3.1 PLS Scheme Using Deep Neural Network Models

DL based PLS scheme for IIoT is proposed by Liao [26] has three lightweight authentication techniques based on DNN, CNN, and Convolutional Preprocessing Neural Network (CPNN). Out of the three authentication mechanisms, CPNN based node authentication has many advantages like a minimal requirement of computational resources, low latency, and light weightiness. Figure 9 depicts the DNN model consists of an input layer, several hidden layers, and an output layer. The CNN based model has two convolutional layers, two pooling layers, and a dense layer in between the input and output layer. The CPNN model is similar to CNN with a brief training span having higher accuracy and authentication. The basic distinction lies in the prior procedure of offline convolution preprocessing. The performance metric used for analysis is the authentication rate and the important finding of this work is the interrelation between the number of hidden layers and authentication accuracy. The accuracy will get increased initially, but it will saturate at one point. The number of hidden layers greater than 5 does not have an impact on the authentication rate of the proposed network.

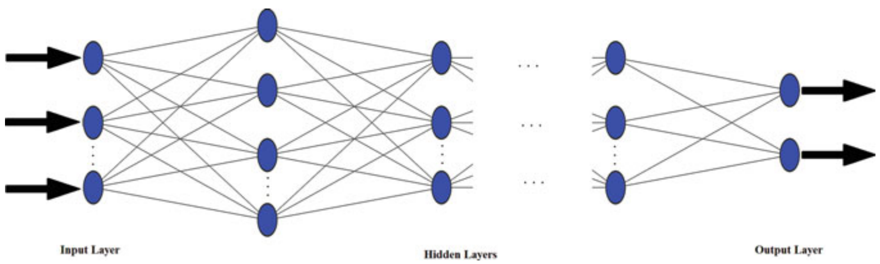
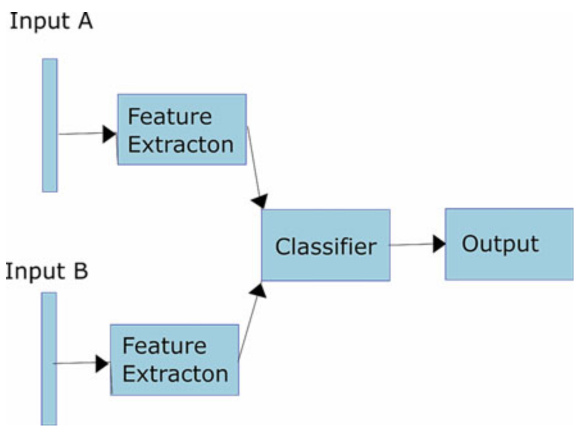


Fig. 9 DNN model for physical layer authentication

3.2 PLS Scheme Using Convolutional Neural Network Models

Abyaneh [27] proposes a DL based physical layer authentication based on the features extracted from the CSI of the legitimated nodes. The proposed architecture namely LocNet consists of dual feature extractors and a final classifier as shown in Fig. 10. The feature extractor is formed by stacking two convolutional layers as shown in Fig. 11, having identical attributes and a final series of dense layers for flattening and compression of data. Batch normalization and Rectified Linear Unit (ReLU) activation functions are used for all the hidden layers. The classifier is a simple MLP model consists of three dense layers having number neurons six, six, and one respectively with activation function ReLU for all the layers except output layers where the sigmoid activation function is used. The authentication methodology explained in [21] can be summarized as in Fig. 12. The major advantage of this method is its robustness towards the rotation of the node's CSI. This scenario is highly probable in IoT, but the proposed system is not resilient to complete change in the environment. Hence, this does not apply to mobile IoT networks. The practical tests are done on different locations and verified the results in two environments, one an apartment and the other one in a garage.

Fig. 10 LocNet architecture



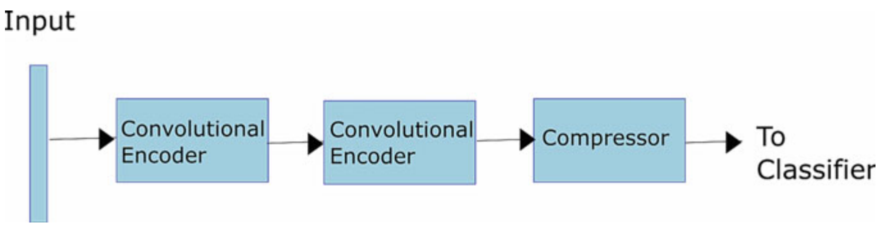


Fig. 11 CNN based architecture of feature extractor

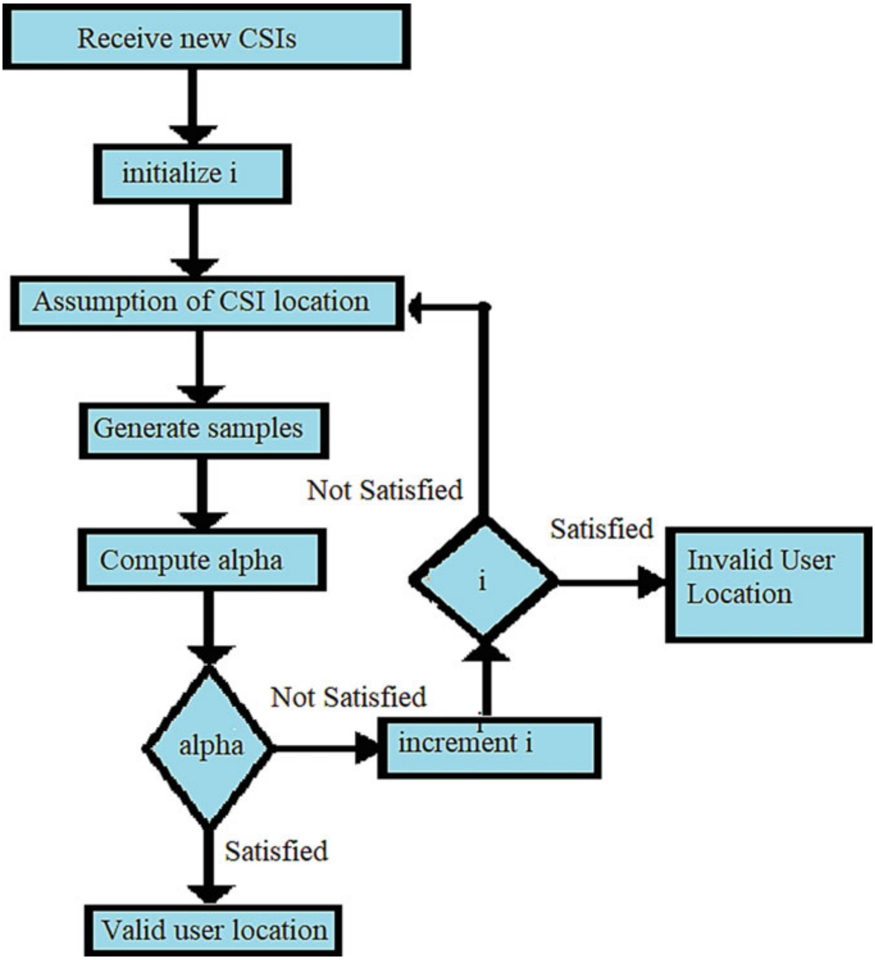


Fig. 12 CNN based PLS algorithm

Qiu et al. [28] proposed a Lightweight Physical Layer Authentication (LPLA) scheme based on the deep learning model for multimedia IoT networks. Here the authors consider the DL model for blind feature extraction from the estimated channel matrix. The CNN is used for taking out the blind features contains two convolutional layers and pooling layers, which is followed by a dense layer. Softmax loss is considered as the target function of the training optimization. ML model uses the multiple physical feature sets and an SVM based classifier for the detection of a spoofing attack. The analysis of detection accuracy considering different physical layer attributes such as Received Signal Strength (RSS), Distance between Adjacent Signals (DAS), and Pearson Correlation Coefficient (PCC) and the combination of these is done. The combination of three features has higher detection accuracy out of

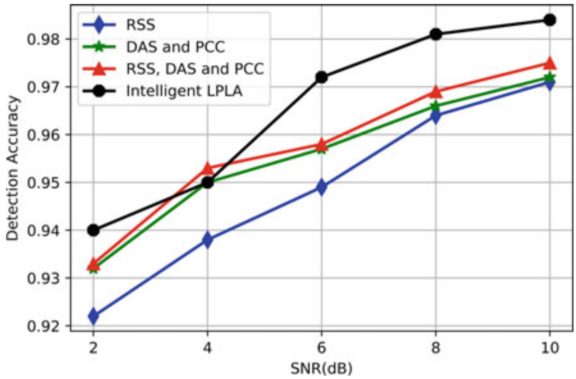


Fig. 13 Comparison of combination of different features in terms of detection accuracy

these three feature sets. The proposed DL based LPLA solutions are compared based on the detection accuracy and the prominence of the DL based method is proven as we can infer from Fig. 13.

Qiu et al. [29] proposed an adaptive neural network based on the CNN model to address spoofing-based attacks in wireless networks as represented in Fig. 14. The dynamic physical characteristic is tracing out using Data-Adaptive Matrix (DAM). The proposed system is modeled on Universal Software Radio Peripheral (USRP) and analyzed considering a conference room scenario.

Baldini et al. [30] proposed a fusion of CNN and Recurrence Plot (RP-CNN) for the physical layer authentication of IoT networks is shown in Fig. 15. Recurrence plots are the 2D representation of repeated return to the phase space states. The results suggest that RP-CNN with threshold 0.01 is having 96.8% accuracy whereas 99.4%, 98.2%, 97.9 for the threshold values of 0.02, 0.06, and 0.07 respectively, where the threshold represents Radiometric Identification (RAI) parameter.

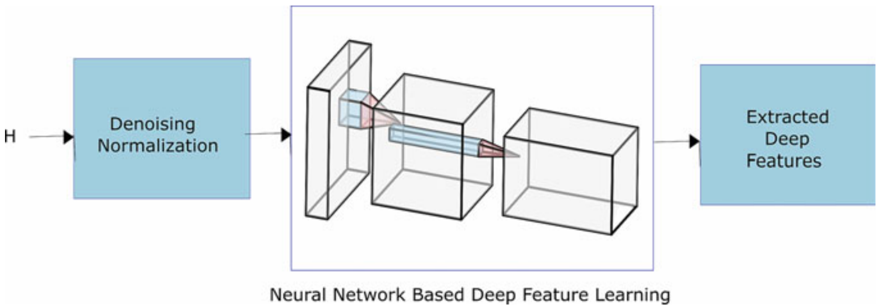


Fig. 14 DNN model for feature extraction

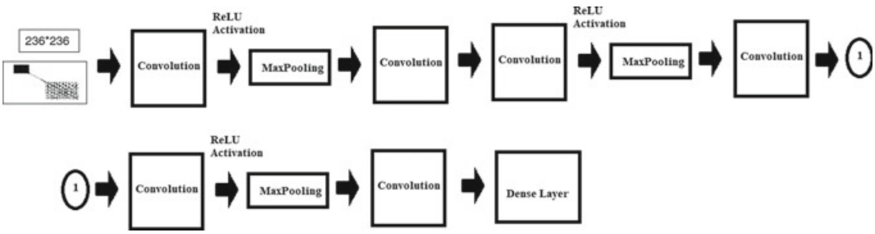
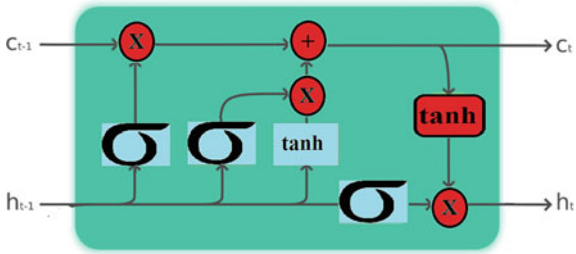


Fig. 15 CNN model in the proposed RP-CNN based physical layer security

3.3 PLS Scheme Using Long Short-Term Memory Networks

Ferdowsi et al. proposed [31] LSTM based PLS for an adaptive authentication to avoid man in the middle attack and data injection attack for the IoT networks. The basic architecture of an LSTM memory unit is depicted in Fig. 16. The LSTM based dynamic watermarking framework is introduced in the work based on the IoT Devices (IoTD) feature set identified by the IoT gateway. Further, for scalability gateway is equipped with a game-theoretic scheme. The flatness of the spectral resource, skewness, kurtosis, and variance is chosen as the features that need to be extracted from the IoTD signals. These feature sets are mapped with a bit sequence of fixed length which is then used for imprinting on the signal as a watermark using the key. The input to the LSTM network is signal and Pseudo Noise (PN) key and the output of the network is the watermarked signal. The proposed architecture is shown in Fig. 17 where the Dynamic Watermarking LSTM (DW-LSTM) extracts the features from the signal using the key. Based on the extracted bits and extracted

Fig. 16 Architecture of a single LSTM unit



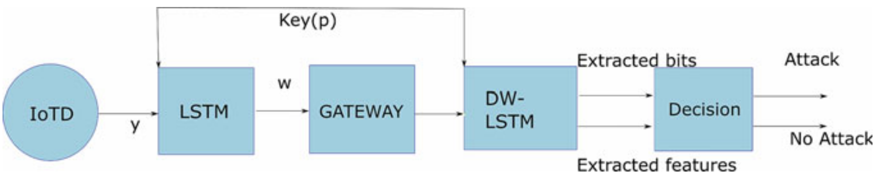


Fig. 17 LSTM based watermark authentication scheme

features the decision on the device's authenticity is taken. The equilibrium based on Mixed Strategy Nash Equilibrium (MSNE) is being derived based on the deep reinforcement algorithm which detects the exposed IoTDs.

3.4 PLS Schemes Based on Recurrent Neural Networks

Way et al. [32] proposed PLS exploiting CSI of the channel for DL aided authentication scheme to detect spoofing attacks. The CNN, RNN, and also a hybrid model of both the above namely Convolutional Recurrent Neural Network (CRNN) are used for authentication purposes. The DL models proposed maps the CSI of the transmitting nodes to the uniqueness of those transmitting nodes, which lies in the interrelation between CSI values of different times for a single terminal. The CNN based model is chosen considering the virtue of unwavering effect on the environmental changes and imperfections. RNN based model considers CSI data as a sequential data model and the dependencies for different frequency components of CSI are exploited. To enable the merits of both the model CRNN model is proposed. The CNN model is formed by establishing the following sequence of neural layers. An input layer, the stack of convolutional layer and pooling layer, followed by a flatten layer and a dense layer connected to the output layer. The output layer provides an authentication decision. The RNN model can seize the contextual information for the attestation of individual nodes' legitimacy. The recurrent layers in the RNN link the preceding states to the present states using feedback loops as shown in Fig. 18. The architecture of the CRNN model is depicted in Fig. 19 contains a multitude of convolutional layers and pooling layer duo followed by recurrent layers. The performance comparison of these models with the benchmark model is carried out by the authors in which 97%, 98.6%, and 99.7% are the accuracy of CNN, RNN, and CRNN respectively. Further, the false alarm rate and misdetection rate data also support the dominance of the CRNN over CNN and RNN.

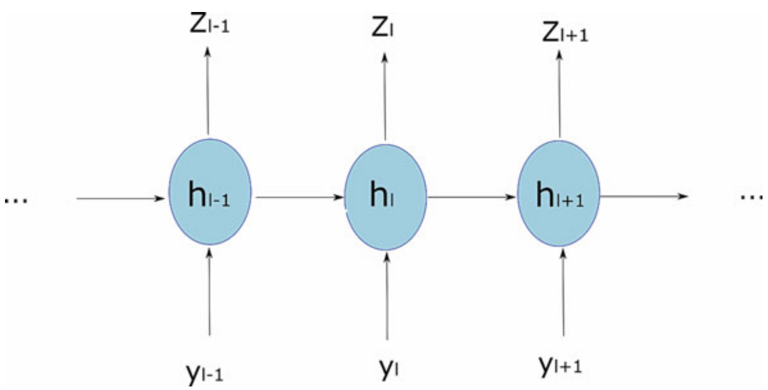


Fig. 18 RNN model for PLS

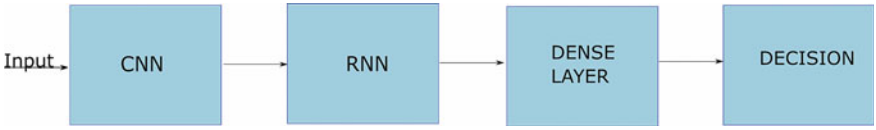


Fig. 19 CRNN architecture for physical layer authentication

4 Fuzzy Systems for Physical Layer Security for IoT

Fuzzy systems are one of the pioneering soft computing techniques where the computing paradigm is knowledge-driven. The knowledge base is updated with fuzzy if-then laws. As represented in Fig. 20. fuzzy systems have a fuzzifier to convert numeric input to fuzzy input, a knowledge base, a fuzzy inference engine, and a de-fuzzifier. Thu fuzzifier fuzzifies the non-fuzzy data which in turn passed to the inference engine for the extraction of fuzzy if-then inferences and stored in the knowledge base. The deffuzzifier is used to deliver numeric output from the fuzzy inference engine.

Fuzzy systems are also having a significant role in providing Physical layer security of IoT. Fang et al. [33] studied the performance of a PLS scheme based on the fuzzy systems. The proposed multidimensional adaptive multi-dimensional PLS exploits the virtue of fuzzy logic. The proposed parametric technique requires a smaller number of samples compared to the non-parametric method. The system model is depicted in Fig. 21 is characterizes the secure transmission of a message between Alice and Bob with a spoofing attack by Eve. Bob needs to authenticate the sender through defective estimates of physical layer attributes like channel gain. The imperfect multi-attributes are considered as the labels for the fuzzy set which considers the transition from authentic to unauthentic a continuous one rather than an abrupt one. This attribute of fuzzy-based transition makes the final authentication more robust and accurate. The multiple physical layer properties are combined using a product inference engine and center average de-fuzzifier. The fuzzy membership function employed for the attribute combination can be Gaussian, Trapezoidal, and Triangular function. The authors considered Gaussian function as the fuzzy

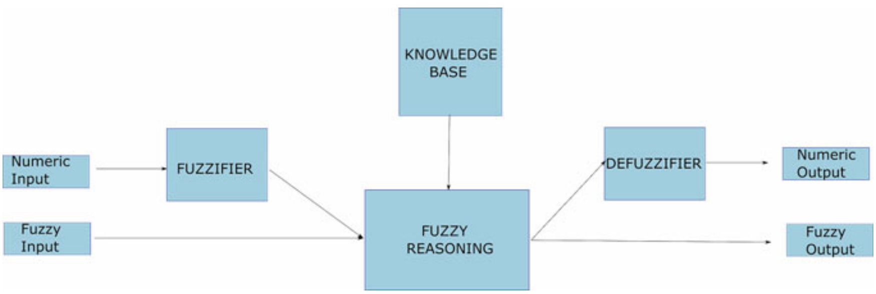


Fig. 20 Architecture of a typical fuzzy system

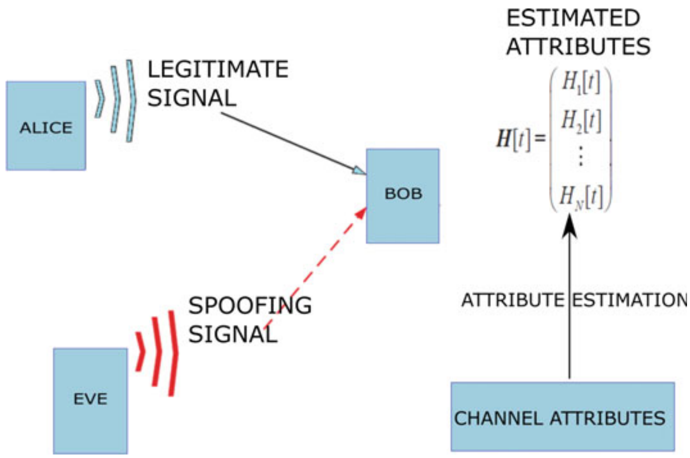


Fig. 21 Physical layer authentication scheme proposed by Fang et al. [25]

membership function with mean μ_n and variance σ_n^2 with distribution

$$u(-1, 1)h_n'(t) = e^{-\left(\frac{(h_n'(t)-\mu_n)^2}{2\sigma_n^2}\right)} \tag{6}$$

Theoretical analysis of authentication performance and empirical validation considering various simulation scenarios are validated. The proposed system is compared with a few state-of-the-art authentication schemes and its prominence. The major advantage of this technique is its adaptability to IoT systems like smart building and smart factories etc. Vishal Sharma et al. [34] proposes Behavior Rule specification-based misbehavior detection for IoT (BRIoT). The PLS verification is carried out using a two-layer fuzzy-based model namely Hierarchical Context-Aware Aspect-Oriented Petri Net (HCAPN) is shown in Fig. 22. A modified Fuzzy C-Mean (FCM) algorithm is used for clustering. The correctness of the behavior rules for IoT devices is verified by analyzing the acquiesce data for the IoTDs by using BRIoT. The results affirm the robustness and effectiveness of the proposed model. Wang et al. [35] proposed a PLS to address spoofing attacks by applying feature extraction and recognition. To address the formulated problem a fuzzy c-mean algorithm is applied. The preprocessing of the physical features by Sparse Representation (SR) is carried out to fortify the features. The performance of the system is analyzed considering Line of Sight (LOS) and Non-Line of Sight (NLOS) scenarios, considering different features like RSS, Channel Impulse Response (CIR), and Multipath Delay (MD).

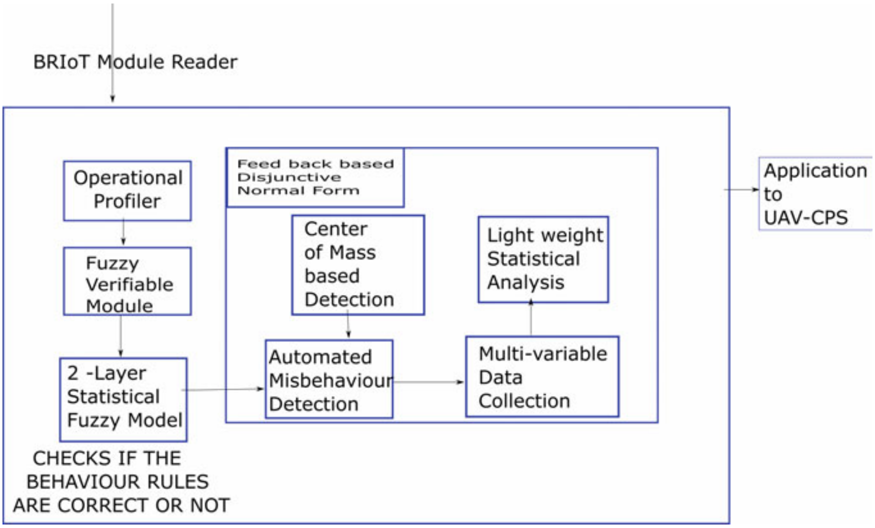


Fig. 22 Design of BRIoT based PLS scheme

5 Genetic Algorithm for Physical Layer Security for IoT

Meta-heuristic algorithms are widely functioning as black-box for the partial solution of optimization problems. The list of meta-heuristic algorithms is very vast. The major contenders are the GA, Particle Swarm Algorithm (PSA), Bat Algorithm (BA), Ant Colony Optimization (ACO), etc. A Multi-Objective Genetic Algorithm (MOGA) based optimization stratagem is put forward by the authors the design parameters such as length, width, and spacing considering the secrecy as objective. The algorithm for the MOGA framework is as shown in Fig. 23. The secrecy capacity of the proposed work is assessed for the varying distance between transmitter and eavesdropper. The secrecy capacity value is a near-optimal improvement. Okati et al. [36] conducted a comparative study of the GA, PSO, Bee algorithm, and Teaching Learning Based Optimization (TLBO) for PLS power allocation. The parameters selected for each algorithm listed in Table 2. The computational resource requirement is listed in Table 3. All the algorithms are applied by considering the scenario of three intermediate corporative nodes and one relay, two cooperative jammers, and an eavesdropper. The GA and TLBO approach found superior to other meta-heuristic techniques. Oliveira et al. [37] proposed Genetic algorithm based PLS, where the IoT network is equipped with the cognitive radio techniques and the security achieved by employing improper Gaussian signals for secondary users. The upper hand of the proposed algorithm is verified by the vigorous analysis of performance by using performance metrics such as secrecy outage probability, effective secure throughput, and secure energy efficiency.

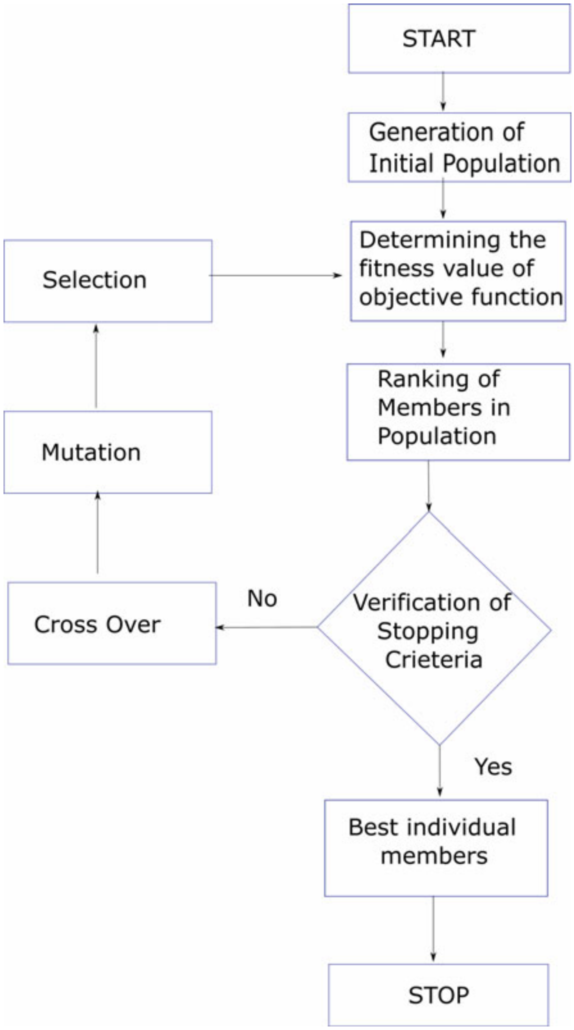


Fig. 23 Multi-objective Genetic Algorithm frame work

Rahman et al. [38] proposed a genetic algorithm based PLS for the cooperative cognitive radio network. The GA based optimization is used for maximizing the secrecy rate of the system considering an eavesdropper in the network. The GA based approach can produce a better secrecy rate with comparatively less computational resources.

Table 2 Selected parameters for meta-heuristic algorithms based PLS [36]

Algorithm	Parameters	Value
GA	Population size	100
	Crossover percentage	0.7
	Mutation percentage	0.3
	Mutation rate	0.1
PSO	Population size	100
	Inertia weight	1
	Inertia weight dumping ratio	0.99
	Personal learning coefficient	1.5
	Global learning coefficient	2
BA	Number of Scopus Bees	100
	Neighborhood radius damp rate	0.95
TLBO	Population size	100
	Teaching factor	1–2

Table 3 CPU Time requirement for the meta-heuristic algorithms based PLS [36]

Algorithms	CPU time (s)
GA	0.77
PSO	1.4
BA	12.8
TLBO	0.67

6 Conclusions

The growing demand for IoT based networks creates the urge for a paradigm shift in the security architecture of IoT. Recent attacks in vulnerable IoT network creates a devastating effect in different arenas where IoT have remarkable role especially in areas like healthcare. The physical layer security is a state-of-the-art security paradigm based on the information-theoretic aspects of the physical layer. The PLS has an indispensable role in providing lightweight and cost-effective security solutions for IoT based network. Various PLS techniques are proposed in the literature. One of the flourishing domains in PLS is soft computing based PLS approaches. Different soft computing approaches come from both knowledge-driven and data-driven techniques that have contributions in felicitating PLS in IoT. One of the major advantages of this type of PLS technique is its inherent ability to adapt to the varying environment. Another advantage of these techniques is its ability to improve its performance after real-time implementation. The major challenge in soft computing based PLS techniques is the need for dedicated hardware. The lightweight and computationally efficient hardware will further elevate the role of soft computing techniques in PLS for IoT.

References

1. Mukherjee, A.: Physical-layer security in the internet of things: sensing and communication confidentiality under resource constraints. Proc. IEEE **103**(10), 1747–1761 (2015). https://doi.org/10.1109/JPROC.2015.2466548
2. Choi, J.: Secure transmissions via compressive sensing in multicarrier systems. IEEE Signal Process. Lett. **23**(10), 1315–1319 (2016). https://doi.org/10.1109/LSP.2016.2595524
3. Dong, L., Han, Z., Petropulu, A.P., Poor, H.V.: Improving wireless physical layer security via cooperating relays. IEEE Trans. Signal Process. **58**(3), 1875–1888 (2010). https://doi.org/10.1109/TSP.2009.2038412
4. Hu, L., et al.: Cooperative jamming for physical layer security enhancement in internet of things. IEEE Internet Things J. **5**(1), 219–228 (2018). https://doi.org/10.1109/JIOT.2017.2778185
5. Zhang, X., McKay, M.R., Zhou, X., Heath, R.W.: Artificial-noise-aided secure multi-antenna transmission with limited feedback. IEEE Trans. Wirel. Commun. **14**(5), 2742–2754 (2015). https://doi.org/10.1109/TWC.2015.2391261
6. Wang, G., Meng, C., Heng, W., Chen, X.: Secrecy energy efficiency optimization in AN-aided distributed antenna systems with energy harvesting. IEEE Access **6**, 32830–32838 (2018). https://doi.org/10.1109/ACCESS.2018.2846689
7. Zhang, J., Marshall, A., Hanzo, L.: Channel-envelope differencing eliminates secret key correlation: LoRa-based key generation in low power wide area networks. IEEE Trans. Veh. Technol. **67**(12), 12462–12466 (2018). https://doi.org/10.1109/TVT.2018.2877201
8. Wang, S., Li, W., Lei, J.: Physical-layer encryption in massive MIMO systems with spatial modulation. China Commun. **15**(10), 159–171 (2018). https://doi.org/10.1109/cc.2018.8485478
9. Benzid, D., Kadoch, M., Cheriet, M.: Raptor code based on punctured LDPC for secrecy in massive MiMo. In: 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, pp. 1884–1889 (2019). https://doi.org/10.1109/iwcmc.2019.8766490
10. Hornansky, M., Zao, J.K.: Customizing short-length LT codes with evolution strategies for video streaming protection. Multimed. Tools Appl. **76**, 15221–15250 (2017). https://doi.org/10.1007/s11042-016-3756-9
11. Rohokale, V.M., Prasad, N.R., Prasad, R.: Reliable and secure cooperative communication for wireless sensor networks making use of cooperative jamming with physical layer security. Wirel. Pers. Commun. **73**, 595–610 (2013). https://doi.org/10.1007/s11277-013-1205-x
12. Chun, C., Kang, J., Kim, I.: Deep learning-based channel estimation for massive MIMO systems. IEEE Wirel. Commun. Lett. **8**(4), 1228–1231 (2019). https://doi.org/10.1109/LWC.2019.2912378
13. Siyad, C.I., Tamilselvan, S., Sneha, V.V.: Frequency domain learning scheme for massive MIMO using deep neural network. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1293–1300 (2020). https://doi.org/10.1109/iciccs48265.2020.9121094
14. Qing, C., Cai, B., Yang, Q., Wang, J., Huang, C.: Deep learning for CSI feedback based on superimposed coding. IEEE Access **7**, 93723–93733 (2019). https://doi.org/10.1109/ACCESS.2019.2928049
15. Siyad, C.I., Tamilselvan, S.: Deep learning enabled physical layer security to combat eavesdropping in massive MIMO networks. In: Smys, S., Senjyu, T., Lafata, P. (eds.) Second International Conference on Computer Networks and Communication Technologies. ICCNCT 2019. Lecture Notes on Data Engineering and Communications Technologies, vol. 44. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37051-0_72
16. Huang, H., Yang, J., Huang, H., Song, Y., Gui, G.: Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system. IEEE Transa. Veh. Technol. **67**(9), 8549–8560 (2018). https://doi.org/10.1109/TVT.2018.2851783

17. Urano, K., Hiroi, K., Yonezawa, T., Kawaguchi, N.: An end-to-end BLE indoor location estimation method using LSTM. In: 2019 Twelfth International Conference on Mobile Computing and Ubiquitous Network (ICMU), Kathmandu, Nepal, 2019, pp. 1–7 (2019). https://doi.org/10.23919/icmu48249.2019.9006638
18. Fathimath Shamna, K., Ismayil Siyad, C., Tamilselven, S., Manoj, M.K.: Deep learning aided NOMA for user fairness in 5G. In: 2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2020, pp. 1–6. https://doi.org/10.1109/icsss49621.2020.9202308
19. Ji, X., Wang, J., Li, Y., Sun, Q., Xu, C.: Modulation recognition in maritime multipath channels: a blind equalization-aided deep learning approach. China Commun. **17**(3), 12–25 (2020). https://doi.org/10.23919/JCC.2020.03.002
20. Li, W., Yi, P., Wu, Y., Pan, L., Li, J.: A new intrusion detection system based on KNN classification algorithm in wireless sensor network. J. Electr. Comput. Eng. **2014**, 240217:1–240217:8 (2014)
21. Senigagliesi, L., Baldi, M., Gambi, E.: Performance of Statistical and Machine Learning Techniques for Physical Layer Authentication (2020). ArXiv, abs/2001.06238
22. Gianmarco, B., Raimondo, G., Gary, S.: Physical Layer authentication and identification of wireless devices using the Synchrosqueezing transform. Appl. Sci. **8**, 2167 (2018)
23. Hoang, T.M., Duong, T., Tuan, H.D., Lambotharan, S., Garcia-Palacios, E., Nguyen, L.D.: Physical Layer Security: Detection of Active Eavesdropping Attacks by Support Vector Machines. arXiv: Signal Processing (2020)
24. Chen, S., Wen, H., Wu, J., Chen, J., Liu, W., Hu, L., Chen, Y.: Physical-layer channel authentication for 5G via machine learning algorithm. Wirel. Commun. Mob. Comput. **2018**, 6039878:1–6039878:10 (2018)
25. Yu, J., Hu, A., Zhou, F., Xing, Y., Yu, Y., Li, G., Peng, L.: Radio frequency fingerprint identification based on denoising autoencoders. In: 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1–6 (2019)
26. Liao, R., Wen, H., Wu, J., Pan, F., Xu, A., Jiang, Y., Xie, F., Cao, M.: Deep-learning-based physical layer authentication for industrial wireless sensor networks. Sensors (Basel, Switzerland) **19** (2019)
27. Abyaneh, A.Y., Pourahmadi, V., Foumani, A.H.: CSI-based authentication: Extracting stable features using deep neural networks. Trans. Emerg. Telecommun. Technol. **31** (2020)
28. Qiu, X., Du, Z., Sun, X.: Artificial intelligence-based security authentication: applications in wireless multimedia networks. IEEE Access **7**, 172004–172011 (2019)
29. Qiu, X., Dai, J., Hayes, M.: A learning approach for physical layer authentication using adaptive neural network. IEEE Access **8**, 26139–26149 (2020). https://doi.org/10.1109/ACCESS.2020.2971260
30. Baldini, G., Giuliani, R., Dimc, F.: Physical layer authentication of Internet of Things wireless devices using convolutional neural networks and recurrence plots. Internet Technol. Lett. **2** (2019)
31. Ferdowsi, A., Saad, W.: Deep learning for signal authentication and security in massive internet-of-things systems. IEEE Trans. Commun. **67**, 1371–1387 (2019)
32. Wang, Q., Li, H., Zhao, D., Chen, Z., Ye, S., Cai, J.: Deep neural networks for CSI-based authentication. IEEE Access **7**, 123026–123034 (2019)
33. Fang, H., Wang, X., Xu, L.: Fuzzy learning for multi-dimensional adaptive physical layer authentication: a compact and robust approach. IEEE Trans. Wirel. Commun. **19**, 5420–5432 (2020)
34. Sharma, V., You, I., Yim, K., Chen, I., Cho, J.: BRIoT: behavior rule specification-based misbehavior detection for IoT-embedded cyber-physical systems. IEEE Access **7**, 118556–118580 (2019). https://doi.org/10.1109/access.2019.2917135
35. Wang, N., Li, W., Jiang, T., Lv, S.: Physical layer spoofing detection based on sparse signal processing and fuzzy recognition. IET Signal Process. **11**(5), 640–646 (2016). https://doi.org/10.1049/iet-spr.2016.0378

36. Okati, N., Mosavi, M.R., Behroozi, H.: A comparison between different meta-heuristic techniques in power allocation for physical layer security. Iran. J. Electr. Electron. Eng. **13**, 310–317 (2017)
37. Oliveira, G., Fernández, E., Mafra, S., Sánchez, S.M., Azurdia-Meza, C.A.: Optimal improper gaussian signaling for physical layer security in cognitive radio networks. Secur. Commun. Netw. **2018**, 9065856:1–9065856:13 (2018)
38. Rahman, M.A., Lee, Y., Koo, I.: Joint relay selection and power allocation through a genetic algorithm for secure cooperative cognitive radio networks. Sensors (Basel, Switzerland), **18**

C. Ismayil Siyad is currently an Assistant Professor with the MEA Engineering College, Kerala, India. He is a research scholar at Pondicherry Engineering College, Puducherry, India. His publications include 5 journals and 6 conferences. He received M.E. in Communication systems. He attended more than 20 workshops and short-term courses. His research interests are Deep learning for 5G Communications, Physical layer Security, and IoT.

Dr. S. Tamilselvan is currently an Associate Professor with the Pondicherry Engineering College, Puducherry, India. His publications include 45 journals and 40 conferences. He received a Ph.D. in Wireless Communication and Networks. He attended more than 50 short term courses, conducted more than 30 short term courses, and delivered more than 10 expert lectures. His distinctions include the member of ISTE and the Board of studies member of Pondicherry University. He worked as a reviewer for many well-reputed journals and conferences. He supervised 7 research scholars. His research interests are Wireless Communication and Networks, VLSI Design, Solid State Devices.

Linear Congruence Generator and Chaos Based Encryption Key Generation for Medical Data Security in IoT Based Health Care System

Anirban Bhowmik, Sunil Karforma, Joydeep Dey, and Arindam Sarkar

Abstract In the healthcare industry privacy and security of patient's information is the most crucial issue at present. Considering current legal regulations, every healthcare organization should impose a prominent security technique to maintain a secure electronic health records system. On the other hand the realization of Internet of Things is the most notable advancement in the field of computer science and electronics. The healthcare services have increased with the help of IoT. Now-a-days security flaws on patient's information are a significant issue in healthcare system. Electronics health record i.e., collection of health related information are sensitive in nature so, it is very significant to impose an advanced security techniques in the system. Here we have focused on security issue like technical safeguards. Our technique proposes a modified logistic map and linear congruence based security model for securing the diagnostics data of patients and an authentication technique. For encryption and decryption two keys are used which are intermediate key, and session key. The modified logistic map is the backbone of our work. This new approach of key generation provides beauty as well as extra robustness in our technique. The different types of experiments and their results conforms that our technique is very secure and efficient for data transmission in medical sector.

Keywords Internet of things · Intermediate key · Session key · Authentication · Chaos theory · Linear congruence

1 Introduction

Healthcare is defined as the act of taking preventative or necessary procedures to improve a person's well-being. This may be done with surgery, medicine, or other

A. Bhowmik (✉) · J. Dey
Department of Computer Science, M.U.C. Women's College, Burdwan, WB, India

S. Karforma
Department of Computer Science, The University of Burdwan, Burdwan, WB, India

A. Sarkar
Department of Computer Science and Electronics, R.K.M. Vidyamandira, Belur Math, WB, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_3

alterations in a person's lifestyle these services are typically provided by hospitals and physicians through a health care system. At present it is seen that the health-care service is necessary for life. The current trends in healthcare industry are the digitization of healthcare workflow and moving to electronics patient records. An e-record of a patient means electronic version of a patient's medical history like demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports etc. These medical data are maintained by the provider over time. The quantity of e-clinical data has increased in huge rate in terms of complexity, diversity and timeliness. The different types of medical information system have improved by using developed mobile Internet and Internet of things. The key feature of different types of medical information system is that the patients can get their treatment from the health care providers directly. On the other hand, patients can also communicate with each other that have the same symptom. They also may form a group for exchanging their illness-related information, treatment experience, diets, medicines and specialist doctor recommendation. Besides, patients may communicate to encourage each other to overcome the disease, regardless of the patients' locations and conditions. Sometimes, self-confidence and friendly environment are more effective than drugs in patients' conditions.

Internet of Things (IoT) can be defined as a set of networking technologies that consist of different appliances, devices and electronic gadgets to interact and communicate among themselves. The World of IoT includes a huge variety of devices like smart phones, personal computers, PDAs, laptops, tablets, and other hand-held embedded devices. The main theme of communication among IoT devices is to provide the facility for processing information in centralized way and delivered to the intended destinations in cost effective way. In future the facility of IoT provides a Superior world for human beings.

IoT in medical Devices: At present different healthcare systems widely use the IoT devices for smooth functioning, monitoring and assessment of patient's conditions and records [1]. PMD (Personal Medical Devices) are small electronic devices used for monitoring the medical condition of a patient. These Personal Medical Devices (PMDs) are of two types according to its location internal (i.e., inside the patient's body) or external (i.e., attach to the patient body externally). PMDs use a wireless interface to perform communication with a base station that is further used to read status of the device, medical reports, and change parameters of the device, or update status on the device. Unfortunately, the most of these devices and applications are not secure for data or information communication. IoT may prone to different types of attacks. The IoT devices are targeted by attackers and intruders and this may risk for patients. Every health care system should be ensured the security of network in order to prevent the privacy of patient from malicious attacks. To strengthen the sensitive information and other types of security, a proactive, preventive approach and measures must be taken by every healthcare organization with attention to future security and privacy needs.

Importance of Security and Privacy Protection in IoT based health care System: Security is a most important issue in any healthcare system. The attackers aim is to steal the information, attack on devices to utilize patient's resources, or may shut down some applications that are monitoring patient's condition. There are many types of attacks on medical devices that include eavesdropping, in which privacy of the patient is leaked, integrity error in which the message is being altered. The IoT based health care system provides huge benefits in society but it is also prone to different types of attacks. These types of attacks mainly cause information leakage and loss of services in communication channel. The IoT consists of different types of devices and platforms with different credentials and each device and platforms need security according to their characteristics. In IoT platform, lot of personal medical information is shared among various types of devices so the privacy of user is a vital part [2, 3] in health care system. Hence a secure cryptosystem is needed for the data or information protection.

The different types of IoT attacks are physical attack, network attack, software attack, and encryption attack. In our paper we have emphasized only on encryption attacks such as side channel attacks, cryptanalysis attacks, and man in the middle attacks. The different algorithms in our cryptosystem which are described above provide strong protection mechanism of different encryption attack [4]. The result section of our study proves the robustness of proposed cryptosystem by using different data, graphs, comparisons etc.

Comparative Analysis Related to IoT Attacks:

The following Table 1 comparative analyses are related to IoT attacks and it also proves the efficiency of our proposed scheme with respect to different types of parameters.

We compare three attacks with the parameters such as damage level, existing proposal and detection chances, vulnerability, etc. in IoT devices. Sinkhole attack is done at network layer and in this attack the routing information is attracted by the node which is near the base station. The worm attack is taken place at the application layer by inserting malicious code and side channel attack is occurred at both application layer and physical layer. In this attack, the attacker uses the side channel information to generate encryption key. All these attacks except side channel attack are active attacks and can modify the information. All these attacks can modify the data or message, drop the packets, steal the private information and encryption key, etc. At the time of computation each IoT device sends some side channel information to crack the encryption key. Many solutions are provided for each attack but they have some limitations. In our paper, we have emphasized only on side channel attacks with respect to healthcare system. Different algorithms defined above are used to protect the leakage of private information and encryption key.

Our proposed algorithm works on application layer. In our proposed algorithm different types of complex and strong mathematical function from Integer theory and linear function are used and as a result the algorithms when running provides minimum side channel information so that attacker cannot guess the encryption key. Our scheme also provides confidentiality and integrity by checking authentication in both sender and recipient side.

Cryptography and Secret Keys: Cryptography deals with art of codifying messages, so that it become unreadable and can be shared secretly over public

Table 1 Different IoT attacks

Classification/parameters	Sinkhole attack	Worm attack	Side-channel attack
OSI layers	Network	Application	Application
Attack type	Active. It provides the wrong information which results in packet Dropping [5]	Active. It modifies the files [6]	Passive. Here the attacker can find Encryption key by using the side-channel information [7]
Attack threat	Availability, confidentiality-as all the data is attracted to the compromised node	Availability, integrity, authenticity-as it can delete, modify the data [6]	Confidentiality, integrity-by using side channel Information it can find the encryption key [7]
Damage level	High-as all data is flowing through compromised node the attacker can do anything with packet	High-as it can delete files, mail documents [6]	High-as the attacker can obtains the secret key without detecting [7]
Prevention	Yes-if node authentication is provided [6]	Yes-by avoiding suspicious sites, files [3]	Yes-by using preventive methods [8]
Attacks based on	Routing [9]	Malicious code	Side channel information [8]
Vulnerability	Node authentication is not provided [6]	Not following security policies	Side-channel information [8]

communication channels. It is study of developing and using different types of encryption and decryption techniques. Here the plain text is converted into cipher text using an encryption algorithm so that hackers cannot read it, but authorized person can access it. The decryption algorithm works in the reverse order and converts the cipher text into plain text. Cryptography is divided into two types such as symmetric key and asymmetric key cryptography with respect to key.

RSA: The RSA algorithm is a suite of cryptographic techniques and especially it is public key encryption technique. It is widely used in security purpose when sensitive data or information are transmitted through Internet. RSA was first publicly described in 1977 by Ron Rivest, Adi Shamir and Leonard Adleman of the Massachusetts Institute of Technology. In RSA cryptography, both the public and the private key are used in encryption and decryption purposes. RSA has become the most widely used asymmetric algorithm because it assures the confidentiality, integrity, authenticity, and non-repudiation of electronic communications and data storage.

Chaotic System: Chaotic systems [10, 11] are basically nonlinear in nature and exhibiting an apparently random behavior for certain range of values of system parameters. However, the solutions or trajectories of the system remain bounded within the phase space. This unstable state is strongly depending on the values of the parameters and on the way the system begins.

Logistic Map in Chaotic System: The logistic map is a well-known one dimensional chaotic map proposed by R.M. May representing an idealized ecological model for describing yearly variation in the population of an insect species. The mathematical formula is defined as: $y_{n+1} = a * y_n(1 - y_n)$, where a ∈ [0, 4] is the control parameter and $y_0 \in [0, 1]$ is the initial condition [10]. The logistic map shows good behavior and is frequently used in many applications for its chaotic nature in specific range. The dynamics of the logistic map is validated using hopf bifurcation diagrams. The logistic map shows chaotic nature for a ∈ [3.57, 4] and slight variations of the initial value produce major differences in the generated random values. This sequence of values is non-periodic and non-conversing in nature.

Linear Congruence: Let f(x) = $a_0x^n + a_1x^{n-1} + \cdots + a_n (n \geq 1)$ be a polynomial with integer coefficient a $a_0, a_1, \ldots, a_n$ with $a_0 \not\equiv 0\ (modm)$. Then f(x) ≡ $0(modm)$ is said to be a polynomial congruence $(modm)$ of degree n. If there exists an integer x_0 such that $f(x_0) \equiv 0(modm)$, then the solution of the congruence is x_0.

Theorem 1 If gcd $(a, m) = 1$, $h\ h\ ax \equiv b(modm)\ h$.

Proof Since $gcd(a, m) = 1$, there exist integers u, v such that $au + mv = 1$.

Therefore $a(bu) + m(bv) = b$. This gives a $(bu) \equiv b(modm)$. This shows that $x = bu$ is a solution of the congruence $ax \equiv b(modm)$. Let x_1 and x_2 be solutions of the congruence $ax \equiv b(modm)$. Then $ax_1 \equiv b(modm)$ and $ax_2 \equiv b(modm)$. This implies $ax_1 \equiv ax_2(modm) \to x_1 \equiv x_2(modm)$, since $gcd(a, m) = 1$. This proves that the congruence has a unique solution.

The concept of the above theorem is used in generation of intermediate key in our proposed technique [12].

Liner and Non-liner functions: A linear function [12] has the form $y = f(x) = ax + b$. A linear function has one independent variable which is x and one dependent variable which is y. b is constant term or y intercept. a is the coefficient of the independent variable. It is also known as slope and it gives the rate of change of the dependent variable. A linear equation is used in our technique in intermediate key generation.

A simple non linear equation is of the form $x^2 + by^2 = c$. A non linear equation looks like a curve when graphed. Its slope value is variable. The degree of a non linear equation is at least 2 or other integer values. The input and output of a non linear system is not directly related. In this article a non linear function is used in encryption process. The aim of the use of non linear function is to create a non linearity in cipher text [12, 13].

Pell's Equation: Pell's equation is a Diophantine equation of the form $x^2 - dy^2 = \pm 1, x, y \in Z$, where d is a given natural number which is not a square [14].

Lemma *For each non-square positive integer d, there are infinitely many positive integers xandy such that* $\left|x - \sqrt{d}y\right| < \frac{1}{y}$.

Theorem (Lagrange) h ,h $x^2 - dy^2 = 1$ h [14].

Here we focus on the integer solutions of Pell's equation. The two integers for a particular d provide a set of numbers for two or more number of d's which is used for key generation in this article.

2 Literature Survey

Shehab et al. [15], provided a comprehensive study on security issues in IoT networks in his work. Various security requirements such as authentication, integrity, confidentiality were discussed in this paper. This paper provides a comparison among different types of attacks, their behavior, and their threat level. These attacks are categorized into four level which are low-level, medium-level, high-level, and extremely high-level attacks and also suggested possible solutions to encounter these attacks.

Yehia et al. [16], surveyed various types of healthcare applications based on wireless medical sensor network (WMSN). IoT environment is suitable to implement these applications. Also, the different types of hybrid security techniques are discussed for handling the security issues of healthcare systems.

Yen and Guo [10] proposed an idea on encryption method called BRIE based on chaotic logistic map. The bit recirculation of pixels is the basic principle of BRIE. It is controlled by a chaotic pseudo random binary sequence. The secret key of BRIE consists of two integers and an initial condition of the logistic map. Further, Yen and Guo [10] also proposed an encryption method called CKBA (Chaotic Key Based Algorithm) in which a binary sequence is generated using a chaotic system.

Chen et al. [11] and Pareek et al. [17] proposed image encryption scheme where an external secret key (as used by for image encryption and by for text ciphers) of 80-bit and two chaotic logistic maps are employed. The initial conditions for the both logistic maps are derived using the external secret key by providing different weightage to its bits.

Abdelaziz et al. [18, 19], has shown the analysis of the security vulnerabilities and the risk factors detected in mobile medical apps. These apps can be categorized into remote monitoring, diagnostic support, treatment support, medical information, education and awareness, and communication and training for healthcare workers. Categories are done according to its risk factor standards. Eight security vulnerabilities and ten risks factors of mobile security project are detected and analyzed by the World Health Organization.

Amin et al. [20, 19] first proposed Medical Information System (TMIS) where novel authentication and key agreement protocol for accessing remote multi-medical server are introduced. However, Amin et al.'s technique [20] was vulnerable to internal attack, replay attack and the man-in-middle attack, impersonation attack and stolen smart card attack [12, 21].

Al Ameen et al. [22, 19] proposed a security and privacy issues in sensor networks application within healthcare perspective. This focuses on the major social implications like secrecy and security and also analyzes these two issues. However, in this paper Clemens Scott Kruse et al. do not provide any protocol or algorithm for security and privacy in healthcare perspective either in same domain or cross domain healthcare platform. This paper does not provide any glance at physical level security and technical level security. It only takes a look at administrative level security and provides some privacy measure without any protocol or algorithm.

Clemens Scott Kruse et al. [9] analyzed and discussed prominent security techniques for healthcare organization which are seeking to adopt a secure electronic health record system. But this work has failed to provide any suitable robust algorithm or technique to prevent information or data breaches. Today's technical safeguard may not be sufficient for next version of attacks. So, it is very important to establish a robust technical safeguard for medical data security.

Xiaoxue Liu et al. [8] proposed a heterogeneous cross-domain AKA protocol with symptoms-matching in TMIS. In CDAKA protocol a patient who is in PKI-domain can communicate remotely with the other patient who is in IBC-domain. This protocol is composed of four phases. The phases are registration phase, login phase, authentication and key agreement phase. Here two patients realize mutual authentication as well as session key establishment but the generation of session key and authentication process is very complex in the sense that the communicating parties have to face four numbers of phases but our technique provides minimum number of phases for communication between patients.

3 Problem Domain

There are different types of challenges exists in IoT based health care system. Different types of attacks or malicious activity may degrade activity of the medical devices and disrupt the communication system of health care. There are two types of attacks namely Active Attack and Passive Attack. An active attack attempts to alter system resources or affect their operation. A passive attack does not affect system resources but attempts to theft vital information from the system. An attack can be occurred by an insider or from outside of the health organization. Based on the target of the attack, there are three types' attacks against the Internet of Things.

- Attacks against a medical devices.
- Attacks against the masters.
- Attacks against the communication between devices and masters.

A common method of attack involves tampering or altering of the messages is the common method of attack. The medical data and information which are transmitted through IoT environment is very sensitive and vital for treatment. So any changes on these data or information causes risk for patients.

4 Our Contributions

From above introduction and literature review part, it is seen that different types of chaotic map is used for encryption decryption purposes. In this article we have developed a cryptosystem based on modified logistic map and linear congruence and this system acts against security breach in communication network. Here we provide some security algorithms that protect different attacks in the communication network of IoT based health care system. We have focused on two basic things related to message communication which are authentication and encryption-decryption. For authentication we have used hash function and a key, named, intermediate key which is generated by using the two public keys of sender and receiver and the concept of linear congruence and Pell's method. For encryption-decryption, two keys (session key and intermediate key) are used one session key and other intermediate key. The session key is generated by using the modified logistic map. The modified logistic map is an important aspect of our cryptosystem. To get the better non linearity in encryption we have used linear and non linear functions. A rigorous frame structure is used for message transmission in wireless network. Different types of information are accumulated in this structure and this information is used for identity verification, key generation and encryption-decryption.

Modified Logistic Map: In this paper we have modified the logistic map of chaotic system. The modified version is $x_{n+1} = r * x_n\left(1 - x_n^2\right)(1 - x_n)$, where r $\in$ [0, 4.3] is the control parameter and $x_0 \in [0, 1.3]$ is the initial condition. This map is chaotic for r $\in$ [3, 4.3]. The details experiments and results are given in result section.

The main objective is to enhance life quality for people or patients who need support of IoT based healthcare by avoiding unnecessary healthcare costs and efforts, and to provide the proper medical support and treatment at the right time.

5 Proposed Work

Methodology: Our model composed of five algorithms which are—(i) Intermediate key generation for authentication and encryption (ii) Session key generation for encryption (iii) Encryption (iv) Authentication realization (v) Decryption process. Our protocol is described below by a compact algorithm with some modules.

ALGORITHM
Input: - plaintext.
Output: - encrypted file.

Methods: -
Step1. Call SSKG ( ) // session key generation using modified logistic map.
Step2. Call INTMKG ( ) // intermediate key generation using linear congruence.
Step3. Call ENCYP ( ) // file encryption process for generate cipher text for secure transmission.
Step4. Call AUTH_CHK ( ) // Authentication check.
Step5. Call DECYP ( ) //decryption process for generate plain text.

All the steps are given below in details. Here public key of patient means the patient-id provided by the hospital or clinic.

5.1 Session Key Generation

Session key is a secret key for symmetric encryption which is used for a particular transaction or session and is valid for a small period of time. In our scheme modified logistic map is used to create session key. At first the user chooses the numbers within the range of control parameter and initial values. The session key length is also by chosen by user.

ALGORITHM-1 (SSKG ())
Input: values within range of control parameter and initial values.
Output: session key.

Methods:
Step1: set x, r, as double and n, i as integer.
Step2: x & r ← get input within range from user.
Step3: n← key length.
Step4:for $i = 0\ to\ (n-1)$ step1
Step5: $x_n = r * x_n(1.0 - x_n^2)(1.0 - x_n)$ // the values of x_n are the required session key.
end for
Step6: End
This session key is used to generate cipher text from plain text and vice versa. This established session key provides confidentiality for subsequent communication and for each session the session key will fresh. This key is send to recipient end through a frame structure which is described below.

5.2 Intermediate Key Generation

In our scheme we have used the intermediate key for authentication check between two patients and also for encryption process. Here intermediate key is generated by using two public keys of two patients. So the intermediate key may vary for different pair of patients and also session. To generate intermediate key we have used the concept of linear congruence in number theory, three variables linear function and pell's formula. The details algorithm is given below.

ALGORITHM-2 (INTMKG ( ))
Input: Two public keys of two patients and four constants.
Output: Intermediate key.

Methods:
Step1. Set d, i, len, fval as integer (as global variable).
Step2. Set imk [], ps[], pr[] as integer array.
Step3. Set fval ← *key_fn()*. {/* the *key_fn() is a module which provides a value in key generation*.*/}
Step4. pr [] ←public key of receiver.
Step5. ps[]← public key of sender.
Step6. len← get_length (receiver's or sender's public key).
Step7. For i=0 to len
Step8. ps[i] ← ps[i] XOR d and pr[i] ← pr[i] XOR d. /* d is a random value between the lowest ascii value of sender and highest ascii value of receiver's public key.
Step9. if (gcd(ps[i], fval)= = 1) then
Step10. imk[i]← call linear_congr (ps[i], pr[i],fval) {/* linear congruence, the equation ps[i]x≡pr[i] (mod fval) provides unique solution (Theorem1) which is stored in imk[] . This is the required intermediate key.*/}
Step11. else imk[i] ← call linear_congr (ps[i]+1,pr[i],fval)
Step12. End if
Step13. End for
Step14. Stop

Sub Procedure:

key_fn ( ): public key of sender, public key of receiver.
Step1. Set n_1, n_2, n_3 n_4 , d_1 , d_2 as integer variable.
Step2. d_1 ← Value chosen by user and d_2← value chosen by user.
Step3. n_1 & n_2← randomChar (public key of sender) such that $(n_1 \sim n_2) \geq d_1$. If it is not possible to get such numbers n_1 and n_2 with said condition then we take n_1 and n_2 such that $(n_1 \sim n_2)$ nearest to d_1.
Step4. n_3 & n_4← randomChar (public key of receiver) such that $(n_3 \sim n_4) \geq d_2$. If it is not possible to get such numbers n_1 and n_2 with said condition then we take n_1 and n_2 such that $(n_1 \sim n_2)$ nearest to d_2.
Step5. We get two numbers for each n_1, n_2, n_3, n_4 respectively using the pell's equation (x^2 -ny^2=1). Thus we get 8 numbers in total.
Step6. The modulus operation is done on 8 numbers by average of (d_1, d_2).
Step7. Variable number of times the shuffle operation is done on 8 numbers.
Step8. First three or four numbers are used for constant terms in a three variable linear function *f(x,y,z)* and again after shuffle operation remaining numbers are used for values of x, y and z in the function. This module returns the functional value which is used for key generation.
Step9. End sub procedure.

This intermediate key is transmitted to the recipient end through secured channel before encryption. In recipient end this key is used for authentication purpose and encryption purpose.

5.3 Encryption Process

In this phase the intermediate key and session key take part in the encryption process. To provide nonlinearity in encryption process we use circular left shift (CLS) operation in encryption process. A two variable non-linear function ($fn\_enc(sessionkey, intermediatekey)$) is used where ascii value of each character of session key and intermediate key and their difference is used to calculate functional value. The functional value is used as the number of times the circular left shift occurs. Science we use a non linear function so the output of the function is not linear with input and this provides an extra non linearity in cipher text. This type of double encryption with session key and intermediate key provides extra robustness in our technique. In both cases XOR [5] operation is done with a CLS operation at end. An example is given below.

Example Let $fn\_enc(x, y) = \frac{x^2}{y} + y^3$ be a non-linear function.

Where y = ascii value of each character of session key, x = ascii value of each character of intermediate key. Now $m = fn\_enc(x, y)\%(ascii\_diff)$, where $ascii\_diff$ = the ascii difference between each character of intermediate key and session key. The functional value m is used as the number of times the circular left shift occurs in each character of partial cipher text to generate final cipher text.

The encryption algorithm is given below.

ALGORITHM-3: (ENCYP ( ))
Input: - plain text, intermediate key, session key.
Output: - encrypted file.

Method: -
Step1. Set m as integer, $plain\_file$ as plain text file and $cipher\_file, cipher\_final$ as cipher text file.
Step2. Set output_file as temporary file.
Step3. if (! eof) then
Step4. output_file= bit_XOROP ($plain\_file, session\ key$)
Step5. cipher_file = bit_XOROP ($output\_file, intermediate\ key$)
Step6. m= fn_enc (session key, intermediate key) /* one example is given above.*/
Step7. Cipher_final= $cipher\_file \ll m$. // circular left shift operation.
Step8. end if
Step9. Stop

5.4 Generation of Authentication Code and Transmission File

After finishing encryption process we have created a rigorous frame structure (transmission file) with four attributes such as Header, Cipher text [2], Tail_msg and Pading using the module AUTH_CHK (). MD5 hash function is used to generate hash value of intermediate key. The module provides a compact fame format which is ready for transmission to the receiver end.

ALGORITHM-4: (AUTHEN_TRANS ( ))
Input: - intermediate key, session key, cipher text and d_1 , d_2 chosen by user.
Output: - transmission file (trans_file).

Methods:
Step1. Set imk [ ], ssk [ ], Padding [ ], Header_msg [ ], Tail_msg[],enc_val[] as integer array.
Step2. Set trans_file as a file. {/* this file is transmitted to the receiver end */}
Step3. imk [ ] ← intermediate key and ssk[ ]← session key.
Step4. Padding [] ← hash_ValOf(imk[]).
Step5.Header_msg[]←hash ($imk[\,] \ll ((d_1 + d_2)\%asciiVal(1st\ char\ of\ imk[]))\ XOR\ ssk$ [ ].
Step6. Tail_msg[ ]← enc_val[ ] {/* enc_val[] contains encrypted value of d_1 *and* d_2.*/}
Step7. trans_file←Call concate_text (Header_msg [ ], cipher text, Tail_msg[], Padding[]).
Step8. Stop.

If the key size is 16 byte then total frame structure for transmission is given below which is created by the function AUTHEN_TRANS ( ).

HEADER_MSG (32 BYTE)	CIPHER TEXT	TAIL_MSG	PADDING (32 BYTE)

The Tail_msg part of transmission file contains encrypted value of two constants (d_1, d_2). This encryption is done by RSA technique. Padding field contains hash value of intermediate key. This padding field is used for authentication purpose. The session key is recovered from header part using CLS operation and the intermediate key.

5.5 Decryption Phase

In decryption phase, at first receiver generates intermediate key by calling IMKG () and then check authentication using hash value of intermediate key from padding field. After finishing authentication check, session key is generated from HEADER_MSG of transmission file and then decryption process is started using two keys. The entire decryption process is given below by a compact algorithm.

ALGORITHM-5: (DECYP ( ))
Input: Intermediate key, cipher text.
Output: plain text.

Methods:
Step1. Set a, n, i as integer and ssk [ ], imk [ ] as integer array.
Step2. Retrieve d_1 and d_2 from TAIL_MSG of transmission file or frame structure.
Step3. imk [ ] ← intermediate key.
Step4. Call AUTHEN_CHK (imk [ ], trans_file).
Step5. if (true) then
Step6. ssk [ ] ← get_SessionKey(Header_msg of trans_file, d_1, d_2). /* session key generation*/
Step7. Decryption process is done with two keys which is reverse of encryption process.
Step8. else Print *Authentication fails.*
end if
Step9. Stop.

6 Result and Discussion

The above algorithms are implemented in latest version of TURBO C interface in a PC of Intel Core i3 (sixth generation or newer) processor, operating system Microsoft Windows 10 × 64, 8 GB RAM and 500 GB internal storage device. In the following sections, simulated results of the proposed scheme are presented. In our experiments, several sizes of different type files are used as plain text.

A good encryption technique should be robust against different types of crypt-analytic, statistical and brute-force attacks. In this section, we have discussed different type's security analysis of like key space analysis, dictionary analysis etc., different types of randomness analysis, sensitivity analysis, statistical analysis and functionality analysis on our proposed encryption scheme [23].

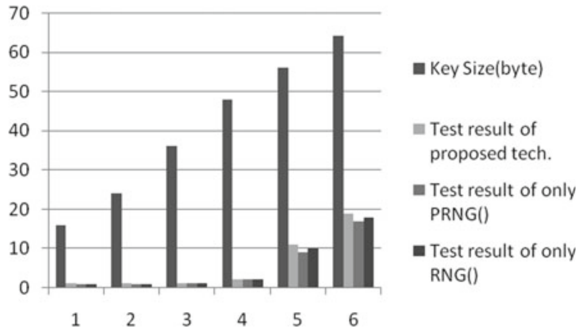
6.1 Randomness Analysis of Key

Randomness means all elements of the sequence are generated independently of each other, and the value of the next element in the sequence cannot be predicted, regardless of how many elements have already been produced. Random and pseudorandom numbers generated for cryptographic applications should be unpredictable (forward and backward). The outputs of a PRNG are deterministic functions of the seed; i.e., all true randomness is depended on seed generation. An RNG uses a non-deterministic source i.e., the entropy source. Here we have used RNG and modified logistic map for generating numbers with true randomness. These random numbers are used for generation of session key. To prove the randomness of our session key we have used serial test [24, 25]. The following statics is used for serial test.

Table 2 Serial test

Key size (byte)	Result of proposed tech.	Result of only PRNG() [24]	Result of only RNG() [25]
16	1.12	1.09	1.08
24	1.2	1	1.01
36	1.26	1.26	1.2
48	2.12	2.11	2.13
56	11	8.98	10
64	19	17	18.03

Fig. 1 Graph of Table 2



$$X = \frac{4}{(n-1)}\left(n_{00}^2 + n_{01}^2 + n_{10}^2 + n_{11}^2\right) - \frac{2}{n}\left(n_0^2 + n_1^2\right) + 1$$

Which approximately follows a χ^2 distribution with 2 degrees of freedom if n $\geq$ 21. The Table 2 [13] is given below.

Observation: For a significance level of $\alpha = 0.05$, the threshold values of X for serial test are 1.12, 1.2, 1.26, 2.12, 11, 19 respectively. Thus the sequence generated by the above algorithm passes serial test. The above Tables 2 and graphs (Fig. 1) prove that our proposed technique is secure against different statistical attacks and differential attacks.

6.2 *Comparative Analysis Between Modified Logistic Map and Standard Logistic Map*

Figure 2 shows the sensitivity analysis on initial condition between logistic map and modified logistic map.

Observations: For a given initial state, the future state can be predicted from a deterministic system. But for chaotic systems, long term prediction of nature of trajectories is impossible. For specific values of parameters, two trajectories, which

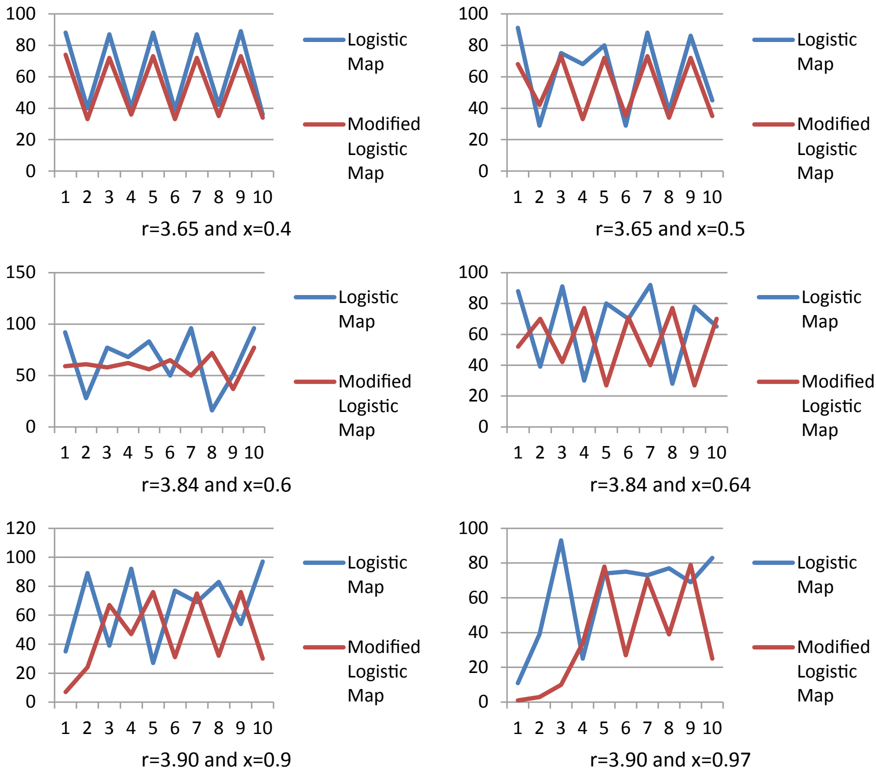


Fig. 2 Compare between modified and standard logistic map

are initially very close, diverge exponentially in a short time. Here initial information about the system is completely lost. From the above figure it is seen that our modified logistic map shows better chaotic nature than standard logistic map within the range. A small change in initial condition with fixed control parameter shows major difference in results as well as graphs.

6.3 Statistical Analysis

Now-a-days different types of statistical attacks and statistical analysis are used by intruders or hackers to analyze the cipher text for decryption. Therefore, an ideal cipher text should be robust against any statistical attacks. To prove the robustness of our proposed encryption scheme, we have performed statistical analysis by calculating the histogram, avalanche effect and randomness by calculating serial test [26].

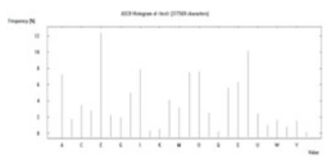
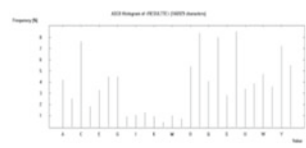
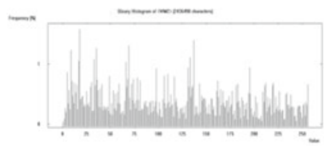
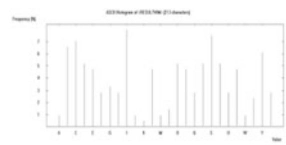
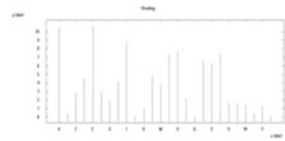
Used Keys:-***encryption@12375 & CrTi182018AB81***		
FILE TYPE	PLAIN TEXT	CIPHER TEXT
TEXT(.txt)		
PDF (.pdf)		
WORD (.docx)		

Fig. 3 Histogram analysis

(a) Histogram Analysis: A text-histogram illustrates [25] how characters in a text are distributed by graphing the number of characters at each level. Here histogram analysis is done on the several encrypted as well as its original text files that have widely different content. We have shown the encrypted files of the original files (plain text) using the secret keys '*ecryption@12375*' and *CrTi182018AB81* (in decimal) in Fig. 3.

Observations: It is clear from Fig. 1 that the histograms of the encrypted files are fairly uniform and significantly different from the respective histograms of the original files (plain text) and hence does not provide any clue to employ any statistical attack on the proposed encryption procedure.

Correlation Analysis: Here we calculate the correlation [27] between ascii difference and total number of changed characters. A secure encryption scheme should transform a text file (.docx, .txt) into a random like encrypted file with low correlation. The formula for Pearson correlation coefficient is given below.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The correlation values are given below through Table 3 [13].

The Pearson correlation coefficient provides the strength and direction of the linear relationship between two variables. From the above Table 1 it is seen that the value of correlation coefficient between X and Y is $-0.426 < 0$. This indicates that

Table 3 Data analysis on correlation

SD of X values	SD of Y values	Correlation coefficient (r)	Coefficient of determination	Significance of test value	Standard error slope values
45.89	200.05	−0.426	18.15%	−1.33	1.394

X values: No. of changed characters, Y values: ascii difference

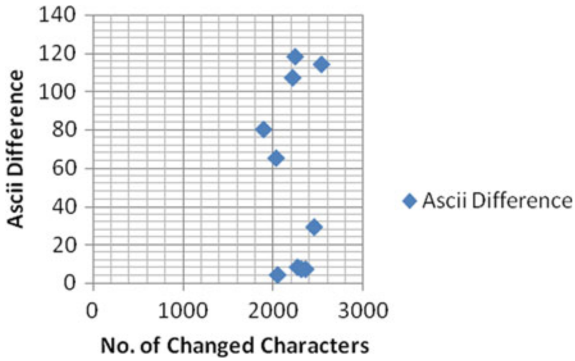


Fig. 4 Graph on correlation analysis

there is a strong negative relationship between the variables or the variables may have a nonlinear relationship. The relationship is negative because, as one variable increases, the other decreases. But from the scatter plot (Fig. 4) [13] it is seen that there is nonlinear relationship exists between two variables (X, Y).

6.4 Key Sensitivity Analysis

An ideal encryption technique should be sensitive with respect to the secret key i.e. a single bit change in the secret key should produce a completely different cipher text. For testing the key sensitivity of the proposed encryption procedure, we have performed the encryption process in the files (.txt) with slight changes in the secret key. The avalanche effect is shown below only for changed session key and with fixed intermediate key. The following Table 4 and graph (Fig. 5) [13] shows the total scenario.

Observation: We have shown the results of some attempts to decrypt an encrypted file with slightly different secret keys than the one used for the encryption of the original file. The above Table 4 shows the added characters, deleted characters and changed characters with slight change (one byte) in session key and the above graph (Fig. 5) also shows the increasing and decreasing performance among added characters, deleted characters and changed characters for one byte change in key. It is clear that the decryption with a slightly different key fails completely and hence the proposed encryption procedure is highly key sensitive.

Table 4 Avalanche effect: change in session key

Key	Ascii difference	Total number of added characters	Total number of deleted characters	Total number of changed characters
encryption@12345	0	3526	3545	2540
eccryption@12345	7	3716	3710	2362
encrydtion @12345	8	3860	3838	2271
encryption @12346	107	3928	3876	2217
encryption #12345	29	3615	3610	2458
encryptinn @12345	80	4262	4245	1892
encryption @1234z	7	3795	3778	2314
encryption @12445	118	3868	3845	2245
encryption @02345	4	4105	4081	2049
Encryption @12345	65	4091	4088	2032

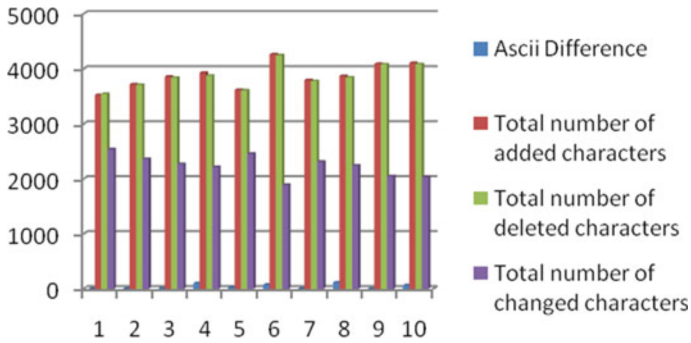


Fig. 5 Graph of avalanche effect

6.5 Security Analysis

Key Space Analysis:

The required property of a good encryption algorithm is the large key space by which algorithm can resist different types of attacks. The total number of different keys in encryption process indicates the size of key space. The brute-force attack is impractical in such crypto systems where key space is large. Now we consider a general case where the size of secret key is k bits i.e., k is the length of key in bits. There are two keys in our proposed scheme, one of which is intermediate key with size k bits and other is session key whose size is k bits also. Now for the intermediate key, the key space is 2^k and for session key, the key space is 2^k and total key space is 2^{2k}. In this large key space platform we discuss following attacks.

Brute-Force Attack:

The large key space makes the brute force attack [28] infeasible. In this attack, attacker tries to translate the cipher text into plain text using every possible key. On average, half of all possible keys are enough for achieving success. Algorithms are known to all in most networking system but brute-force attack will impossible if the algorithm uses large key space. At present the fastest super computer is Tianhe-2 having speeded 200 petaflops i.e., 200×10^{15} floating point operations per second. Let us consider each trial requires 2000 FLOPS to complete one check. So number of trials complete per second is: 100×10^{12}. The number of second in a year is:$365 * 24 * 60 * 60 = 3{,}153{,}600\,\mathrm{s}$.

Now from the above key space the formula for break the key is $2^{2k}/\left(100 * 10^{12} * 3{,}153{,}600\right) = Y$, Y denotes number of years. So if k increases then Y increases ($k \infty Y$) thus for large key length it is difficult to break the key. A cipher text with such a long key space is sufficient for reliable practical use. This proves that a key with large enough is sufficient enough to overcome the brute force attack [29].

Dictionary Attack:

Dictionary attack [30] may uncover available list of words or using this attack password can be found from online. Thus the attacker may get password using dictionary attack. The traditional dictionaries are not only used to find password but also on-line dictionaries of words from foreign languages, or on specialized topics such as music, film, sports etc. are used. For repeated use of these words in encryption process an opponent may create, an "encrypted" (hashed) list of dictionary or high-probability passwords. This dictionary may be used by attacker in guessing right encryption key for decryption. Dictionary attacks are more efficient than a brute force attack because it cannot try nearly as many combinations and if the key is not contained in the dictionary, it will never successfully find it.

In our proposed methodology, the above different algorithms has used random number generation functions, linear congruential function etc. and as a result the intermediate key or session key generated in this way not only contains English words or variations or phrases but also contains different ascii characters, numbers, special characters. This would exhaust attacker's dictionary without a positive match.

Impersonation Attacks: Impersonation [31] means an act of pretending to be another person for purpose of fraud. This type of attack involves message or email that seems to come from a trusted one. If an attacker is intended to get other user's information then the attacker sends a message or email to another user and forge valid information. An impersonation attack can be done when an attacker can successfully forge valid data.

It is required a strong security policies and vigilance on communication media to stop impersonation attack. Here we have used intermediate key for message authentication and RSA technique for user authentication. After authentication we can generate session key which is used for encryption. Thus our technique provides security against impersonation attack.

Replay Attacks: Replay attack [32] is one type of network attack in which an attacker detects a data transmission and it has delayed or repeated fraudulently. As a result attacker can gain access to a network, gain vital information from other or complete a duplicate transaction. An attacker may launch a replay attack to delay or even stop the response to any request message. To defend against replay attack in this article the concept of session key is used and in every session the session key is fresh. Thus when attacker retransmit or has delayed to send message then the previous key may not work.

Tampering Attacks: An attacker tampers [19] other users' information or data using tampering attack. An attacker may launch a tampering attack to a smart system if it intends to change data illegally.

In our proposed methodology, we have introduced a rigorous frame structure with multiple attributes. Among the multiple attributes the cipher text is one so it is hard to detect the range of cipher text. Thus the attacker cannot launch tampering attack easily.

6.6 Functionality Analysis of Encryption Technique

In this section the functionality of our scheme is done by comparing our proposed technology with different standard cryptographic algorithms, different existing schemes and different IoT attacks.

The following Table 5 [13] show comparison among different standard algorithms and our proposed scheme.

The above comparative study shows strength, and acceptance of our scheme. The proposed protocol provides more functionality such as strong user authentication, mutual authentication between the two patients, it establishes a secure session key for the user i.e., patients and these are the paramount requirements for wireless healthcare applications. It is worth notice that our proposed protocol provides indispensable security features. The following Table 6 provides the pertinence of our methodology compared to other existing techniques.

6.7 Significance of Authentication in Our Proposed Scheme

The identity proof is done by Authentication mechanisms [33]. The authentication process ensures the originality of document that is the document is coming from right person or not. In secure system mainly in medical IoT domain the user must identify him/her, and then the system will authenticate the identity before using the system because without proper authentication medical data transmission may occur severe damage in patient party. The authentication process can be professionally seen as: (1) SMS based authentication (2) Intermediate key based authentication (3) Public key authentication. The intermediate key based authentication and public key based authenticity are used in our proposed system, the user shares a single session key with an algorithm named, RSA which provides public key authentication. The hash

Table 5 Comparison between proposed algorithm and standard algorithms

Algorithms	Important features	Important features of our proposed algorithm
IDEA	(i) IDEA encrypts 64-bit plaintext to 64-bit cipher text blocks, using a 128-bit input key (ii) It uses both confusion and diffusion technique (iii) A dominant design concept in IDEA is mixing operations from three different algebraic groups of 2^n elements (iv) The security of IDEA currently seams that it is bounded only by the weaknesses arising from the relatively small (compared to its key length) block length of 64 bits [7, 24]	(i) Our technique encrypts plaintext to cipher text bit by bit, using m-bit input key (ii) The main design concept of our technique (a) intermediate key generation using the concept of linear congruence. (b) Session key generation using modified logistic map. (c) Circular left shift is used to produce non linearity in encryption process
RC5	(i) The RC5 block cipher maintains word-oriented architecture for variable word sizes w = 16, 32, or 64 bits (ii) The number of rounds r and the length of key byte is b byte which is variable (iii) For encryption, there are two steps in each round, (a) bit-wise XOR operation, (b) circular left shift. (c) Addition with the next sub key [27, 25]	(i) Our proposed scheme is stream cipher based. Here two keys are used for encryption/decryption (ii) Key length is variable (iii) For encryption, there are two steps-(a) bit-wise XOR operation (b) Circular left shift with a linear function. It provides number of times CLS occurs
BLOWFISH	(i) This technique is based on stream cipher. It uses addition, XOR operation for encryption (ii) It has a variable key length up to a maximum of 448 bits long which ensures security (iii) Blowfish suits applications where the key remains constant for a long time and it is not suitable for packet switching [7, 27]	(i) Our scheme is also based on steam cipher. It uses XOR, CLS operations to impose more non linearity in cipher text (ii) The use of double keys and one of this key is changeable by nature which provides robustness in our technique (iii) Suitable for packet switching

(continued)

Table 5 (continued)

Algorithms	Important features	Important features of our proposed algorithm
DES	(i) Linear cryptanalysis provides the most powerful attack on DES to date where enormous number of known plain text pairs is feasible (ii) Differential cryptanalysis is one of the most general cryptanalytic tools to date against modern iterated block ciphers, including DES. It is primarily a chosen-plaintext attack (iii) Storage complexity, both linear and differential cryptanalysis requires only negligible storage (iv) Due to its short key size, the DES algorithm is now considered insecure and should not be used However, a strengthened version of DES called Triple-DES is used [24, 25]	(i) Our algorithm is based on stream cipher with two keys one is session key which is changeable in nature. So it protects linear cryptanalysis as well as differential cryptanalysis (ii) Our algorithm takes negligible storage for linear and differential cryptanalysis (iii) Our algorithm is secure with respect to key size, because we have used two keys with variable length

Table 6 Comparison between some existing algorithm and proposed algorithm

Schemes→ Security properties↓	Ref. [33]	Ref. [16]	Ref. [20]	Ref. [34]	Ref. [9]	Ref. [26]	Proposed technique
Confidentiality	Yes	No	No	No	Yes	Yes	Yes
Integrity	No	No	No	No	Yes	Yes	Yes
Authenticity (message authentication and user authentication)	Yes	No	No	Yes	No	Yes	Message authentication using intermediate key and user authentication by RSA
Privacy Protection	No	Yes	No	Yes	Yes	Yes	Yes
Defend against Man-in—middle attack	No	No	No	No	No	No	Yes
Defend against replay attack	No	No	No	No	No	No	Yes
Vulnerability	No	No	No	Yes	Yes	Yes	Yes
Defend against Impersonation attack	No	Yes	No	Yes	Yes	No	Yes
Cryptanalysis (linear and differential)	No	No	No	Yes	Yes	Yes	Yes
Session key establishment	No	No	No	No	Yes	Yes	Yes
Secure against Information-leakage attack	Yes	No	No	Yes	Yes	Yes	Yes

value of amalgamation of two keys i.e., 1st part of intermediate key and 2nd part of session key is concatenated with the plain text. In recipient side, the receiver gets session key using RSA technique and generates intermediate key using Algorithm 2 and after checking the hash value using Algorithm-4 of our scheme, the sender decides whether the message has come from right person or not.

7 Conclusion

There are different types of challenges in health care system at present time. Many possibilities are growing using medical information technologies, but it also pose challenges. Trends in health sensing, application of IoT in healthcare system show interesting new developments. They can enhance and improve healthcare abilities, boost preventive care and foster collaborative healthcare. An integral approach is necessary for healthcare system which abide important aspects such as security and privacy protection. In this paper, we have developed an cryptosystem where an encryption technique is presented based on intermediate key and session key. The session key generation is done by using public key of sender (patient or hospital staff) and receiver (patient or hospital staff) and concept of modified logistic map, linear congruence and linear function. To provide the more non linearity in cipher text we have used circular left shift operation with a non linear function in the encryption technique. Our technique also provides the authentications which enrich the robustness as well as beauty of encryption technique. Different Experimental results prove the feasibility and efficiency of the proposed scheme. Comparative analysis among proposed technique and standard techniques, exhaustive key search analysis, cryptanalysis shows the acceptability of our technique. To the best of our knowledge our proposed technique is the simplest one having minimal computational overhead during encryption and decryption.

Conflict of Interest There is no conflict of interest.

References

1. Agrawal, A., Gorbunov, S., Vaikuntanathan, V., Wee, H.: Functional encryption: new perspectives and lower bounds. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013, Part II. LNCS, vol. 8043, pp. 500–518. Springer, Heidelberg (2013)
2. Joshi, C., Singh, U.K.: A review on taxonomies of attacks and vulnerability in computer and network system. Int. J. Adv. Res. Comput. Sci. Softw. Eng. (IJRCSSE) **5**(1), 742–747 (2015)
3. Ghansela, S.: Network security: attacks, tools and techniques. IJARCSSE **3**(6) (2013)
4. Kruse, C.S., Smith, B., Vanderlinden, H., et al.: Security techniques for the electronic health records. J. Med. Syst. **41**, 127 (2017). https://doi.org/10.1007/s10916-017-0778-4
5. Chaudhry, S.A., et al.: An improved and provably secure privacy preserving authentication protocol for SIP. Peer-to-Peer Netw. Appl. **10**(1), 1–15 (2017)
6. Stinson, D.: Cryptography: Theory and Practice, 3rd ed. Chapman & Hall/CRC (2006)
7. Das, A., Veni Madhavan, C.E.: Public-Key Cryptography: Theory and Practice. Pearson Education (in press)
8. Da Xu, L., He, W., Li, S.: Internet of things in industries: a survey. IEEE Trans. Industr. Inf. **10**(4), 2233–2243 (2014)
9. Abomhara, M., Køien, G.M.: Security and privacy in the internet of things: Current status and open issues. In: International Conference on. Privacy and Security in Mobile Systems (PRISMS), pp. 1–8. IEEE (2014)
10. Yen, J.C., Guo, J.I.: A new chaotic key based design for image encryption and decryption. In: Proceedings of the IEEE International Symposium Circuits and Systems, vol. 4, pp. 49–52 (2000)
11. Chen, G., Mao, Y., Chui, C.K.: A symmetric image encryption based on 3D chaotic maps. Chaos Solitons Fractals **21**, 749–761 (2004)
12. Chakravorty, J.G., Ghosh, P.R.: Advanced Higher Algebra. U.N. Dhur and Sons Private Ltd. (2018). ISBN 978-3- 80673-67-7
13. Bhowmik, A., Karforma, S., Dey, J., Sarkar, A.: Approximation algorithm and linear congruence: a state-of-art approach in information security issues towards internet of vehicles. In: Gupta, N., Prakash, A., Tripathi, R. (eds.) Internet of Vehicles and its Applications in Autonomous Driving. Unmanned System Technologies. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-46335-9_10
14. Shockley, J.E.: Introduction to Number Theory. Hol, Rinehart and Winstont, New York (1967)

15. Boneh, D., Franklin, M.: Identity-based encryption from the Weil pairing. In: Kilian, J. (ed.) CRYPTO2001. LNCS, vol. 2139, pp. 213–229. Springer, Heidelberg (2001)
16. Maia, P., Batista, T., Cavalcante, E., Baffa, A., Delicato, F.C., Pires, P.F., Zomaya, A.: A web platform for interconnecting body sensors and improving health care. Procedia Comput. Sci. **40**, 135–142 (2014). https://doi.org/10.1016/j.procs.2014.10.041
17. Pareek, N.K., Patidar, V., Sud, K.K.: Cryptography using multiple one dimensional chaotic maps. Commun. Nonlinear Sci. Numer. Simul. **10**(7), 715–723 (2005)
18. Oladipupo, E.T., Alade, O.A.: An approach to improve data security using modified XOR encryption algorithm. Int. J. Core Res. Commun. Eng. **1**(2) (2014)
19. Al Ameen, M., Liu, J., Kwak, K.: Security and privacy issues in wireless sensor networks for healthcare applications. J. Med. Syst. **36**, 93–101 (2012). https://doi.org/10.1007/s10916-010-9449-4
20. Olivier, F., Carlos, G., Florent, N.: New security architecture for IoT network. Procedia Comput. Sci. **52**, 1028–1033 (2015). https://doi.org/10.1016/j.procs.2015.05.099
21. Yaling, H.: Research on computer network security analysis model. Comput. CD Softw. Appl. (z), 1–152 (2013)
22. Tarouco, L.M.R., Bertholdo, L.M., Granville, L.Z., Arbiza, L.M.R., Carbone, F., Marotta, M., de Santanna, J.J.C.: Internet of things in healthcare: interoperatibility and security issues. In: IEEE International Conference on Communications (ICC), pp. 6121–6125. IEEE (2012)
23. Zaidan, B., Zaidan, A., Al-Frajat, A., Jalab, H.: On the differences between hiding information and cryptography techniques: an overview. J. Appl. Sci. **10**, 1650–1655 (2010)
24. Stallings, W.: Cryptography and Network Security: Principles and Practice, 3rd ed. Prentice Hall (2003)
25. Kak, A.: Lecture Notes on Computer and Network Security. Purdue University (2015). [Online] https://engineering.purdue.edu/kak/compsec/Lectures.html
26. Liu, X., Ma, W.: CDAKA: a provably-secure heterogeneous cross-domain authenticated key agreement protocol with symptoms-matching in TMIS. J. Med. Syst. **42**, 135 (2018). https://doi.org/10.1007/s10916-018-0985-7
27. Kahate, Atul: cryptography and network security. Tata McGraw-Hill publishing company, New Delhi (2008)
28. Chuang, M.-C., Chen, M.C.: An anonymous multi-server authenticated key agreement scheme based on trust computing using smart cards and biometrics. Expert Syst. Appl. **41**(4), 1411–1418 (2014)
29. Mehlhorn, K., Sanders, P.: Algorithms and Data Structure. Springer, Berlin, Heidelberg (2008). Online ISBN 978-3-540-77978-3
30. Kumar, J.S., Patel, D.R.: A survey on internet of things: security and privacy issues. Int. J. Comput. Appl. **90**(11) (2014)
31. Hwang, Y.H.: Iot security & privacy: threats and challenges. In: Proceedings of the 1st ACM Workshop on IoT Privacy, Trust and Security, pp. 1–1. ACM (2015)
32. Hossain, M.M., Fotouhi, M., Hasan, R.: Towards an analysis of security issues, challenges, and open problems in the Internet of things. In: 2015 IEEE World Congress on Services (SERVICES), pp. 21–28. IEEE (2015)
33. Kumar, V.: Ontology based public healthcare system in internet of things (IoT). Procedia Comput. Sci. **50**, 99–102 (2015). https://doi.org/10.1016/j.procs.2015.04.067
34. Mohan, A.: Cyber security for personal medical devices internet of things. In: 2014 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 372–374. IEEE (2014)

Anirban Bhowmik completed Bachelor of Science (Mathematics) from Bolpur College, Bolpur, West Bengal, India and Master of Computer Application from the University of Burdwan in year 2008. He is working as an Assistant Professor in Department of Computer Application at Cyber Research & Training Institute, Burdwan, West Bengal, India since 2008. He has published ten conferences papers and it's also available online. His main research work focuses on Cryptography and Soft Computing. He has 10 years of teaching experience at UG level.

Sunil Karforma has completed his Bachelors in Computer Science & Engineering, and his Masters in Computer Science & Engineering, from Jadavpur University. He received his Ph.D. in Computer Science, and is presently Professor & Head of the Dept. of Computer Science at the University of Burdwan. His research interests include Network Security, E-Commerce, and Bioinformatics. He has published numerous papers in both national as well as international journals and conferences.

Joydeep Dey pursed Bachelor of Computer Application (Honours) from Cyber Research & Training Institute, Burdwan, India in 2007 and Master of Computer Application from the University of Burdwan in year 2011. He is working as Lecturer in Department of Computer Sciences at M.U.C. Women's College, Burdwan West Bengal, India since 2011. He has published two conferences papers and journals and it's also available online. His main research work focuses on Cryptography and Computational Intelligence. He has 8 years and 0.5 years of teaching experience at UG and PG level respectively.

Dr. Arindam Sarkar is currently serving the Department of Computer Science & Electronics, Ramakrishna Mission Vidyamandira, Belur Math as an Astt. Professor. He has completed M.C.A. degree from VISVA BHARATI, Santiniketan and he is University First Class First. He has completed his M.Tech. in CSE degree from University of Kalyani, India and also secured University First Class First. Dr. Sarkar has completed Ph.D. (Engineering) from University of Kalyani under the INSPIRE Fellowship Scheme of DST, New Delhi. He has secured 2nd Rand in the West Bengal CSC examination. He has more than 55 International Journal and Conference publications.

Content Based Video Retrieval—Methods, Techniques and Applications

Reddy Mounika Bommisetty, P. Palanisamy, and Ashish Khare

Abstract Videos are rich information sources than individual images, they are considered as most influential communication media compared to others. The amount of video data produced and dispensed are growing exponentially day by day with the availability of electronic media such as smart phones, handicams etc. and broadband services at cheaper rates, as well as easy accessibility of those media in the market. Video data storage and access founds its applications in different fields such as digital libraries, video on demand, entertainment etc. and these applications are popular and needs regular access of videos from the libraries. All the above said compound reasons demanded the need of development of efficient video management and retrieval systems which can efficiently retrieve videos similar to the query as well as with a less response time. Video retrieval is made possible by searching of the desired video through a user demanded query. The user inputted query may be in the form of representative keywords or a single image or group of images. The video retrieval systems are classified as text based or content based, according to the query inputted by the user. In a text based video retrieval system query is in the form of representative keywords and the database videos are tagged with appropriate text. An example of concept based search and retrieval system is YouTube. The principal drawback in concept based system is mapping of high level or rich semantics to low level features, which is known as semantic gap. Another drawback in concept based video retrieval systems is intention gap, which denotes gap between query at querying time and intention of the search. Several researchers found content based video retrieval (CBVR) system as solution to the drawbacks of a concept based video retrieval system. The main objective this chapter is to provide comprehensive

R. M. Bommisetty (✉) · P. Palanisamy
Department of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, Tiruchirappalli, India
e-mail: mounika@nitt.edu

P. Palanisamy
e-mail: palan@nitt.edu

A. Khare
Department of Electronics and Communication Engineering, University of Allahabad, Prayagraj, India
e-mail: ashishkhare@hotmail.com

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_4

outlook on content based video retrieval (CBVR) system and its recent developments and a new content based video retrieval system that is going to be developed by feature fusion. The generalized algorithm of CBVR and its individual stages such as keyframe extraction and feature extraction also will be described elaborately. This chapter focuses on a brief overview of CBVR, keyframe extraction, feature extraction and feature fusion.

Keywords Content based video retrieval (CBVR) · Keyframe extraction · Feature extraction · Feature fusion · Semantic gap · Intention gap

1 Introduction

Video is considered as rich powerful and effective way of communication source among all the other communication resources such as text, audio, image etc. In this technological era vast availability of high quality video capturing devices at cheaper rates and internet services even in remote locations made several videos stored on one's personal devices. At one point of time manual search of desired video among the several stored videos becomes a tedious task. Furthermore, cheaper storage devices and high speed web services made lot of video data stored and shared on web. Video is quite popular since it embeds textual, visual and audio embedded in it through which we even convey our emotions along with the information. Different category videos such as sports, e-lectures, traffic, entertainment, surveillance etc. owes their own different in built characteristics. For example, traffic videos contain large movement of vehicles, e-lecture videos contains a stationary background with little bit foreground object movement. Foreground object movement is very less e-lecture videos compared to traffic videos. Based on visual information human can easily understand and interprets about the video. However, for a computer organizing or arranging such all categories of several videos and retrieving those videos later for usage is not such an easy task and involves group of several complex tasks such as spatial/temporal segmentation [1–6], feature extraction [7–10] etc. A video retrieval system is designed for storing, organizing and retrieving videos.

The goal of video retrieval system is to retrieve videos similar to the user defined query. Similarity can estimated through feature extraction and matching. Features can be low level features or high level semantics [11–13] describing the video content. Video retrieval algorithms accepts query either in form of text [11] or sample image or sequence of sample images [14]. Video retrieval algorithms based on text is popular. An example of text based retrieval system is YouTube. In a text based video retrieval system, user inputs query describing about features of video. The algorithm searches through the database and retrieves the videos which are tagged with the user described bag-of-words [15]. The principal drawback of text based retrieval system is the presence of a huge semantic gap [16] in mapping the rich user defined semantics with low level features of the video and thus results a lot of undesired content. Manually annotating all the videos in the database and designing proper keywords to describe the video content is a complex tedious task. Another drawback is intention gap [17] between the query at querying time and intention of the search.

To overcome drawbacks of text based retrieval systems researchers developed Content based Video retrieval system (CBVR) [14] which does not need any manual tagging of keywords with the database videos and retrieval accuracy also improved more compared to text based video retrieval systems. The goal of CBVR is to retrieve videos similar to the query on the basis of visual features present in the video. CBVR systems accept query in form of a sample image or video clip. CBVR extracts features of user inputted query and those features are matched with the features of database videos for retrieving videos similar to the query. Figure 1 shows general CBVR framework. CBVR systems do not require manual tagging and retrieves visually similar videos and most of the CBVR techniques employ keyframe extraction techniques [18–23] for representing video with less number of frames which reduces computational complexity and time.

The above mentioned benefits of CBVR motivated researchers to further explore in this field to improve accuracy even for complex videos.

The rest of the chapter is organized as follows. Section 2 discusses some of the past works in the domain of content based video retrieval. Section 3 discusses the proposed content based video retrieval and its individual stages along with the experimental results. Section 4 discusses few applications of the domain content based video retrieval. Finally, Sect. 5 concludes the chapter.

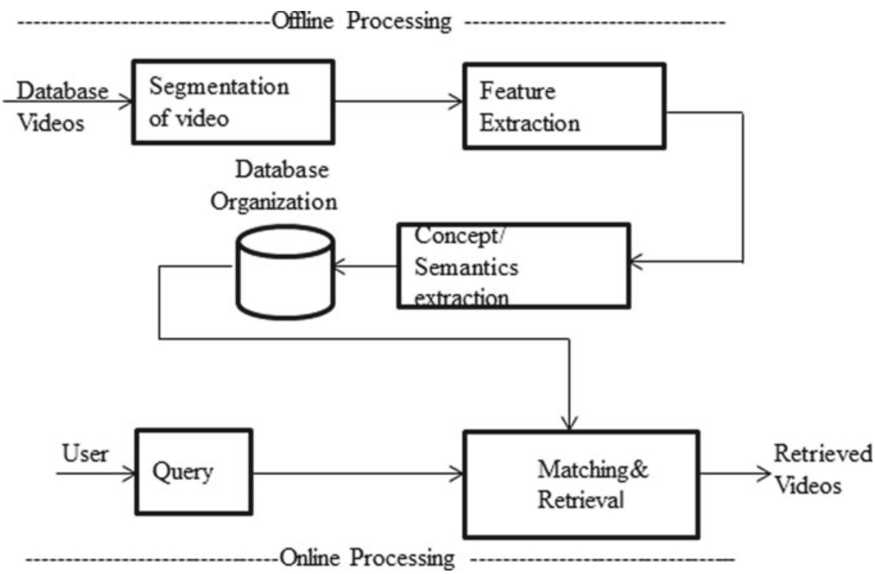


Fig. 1 Content based video retrieval system

2 State-of-the-Art Techniques

Developments in multimedia technology brought vast usage of video data as information or communication source. During recent days, smart phones equipped with high resolution camera are easily available in the market with affordable prices. Technological innovations made high speed web services available to people in cheaper rates and even accessible through smartphones with the same high speed. Several computer vision applications such as video on demand, e-learning, biomedical and surveillance applications as well as availability of several videos on internet made researchers attracted towards video management retrieval. Lokoč et al. [24] presented good review on a class of interactive video retrieval scenarios and their evaluation methods to analyse performance of new interactive video retrieval approaches. Dong et al. [25] presented video retrieval technique in which a text based query is inputted by user and the algorithm searches for unlabelled videos which are tagged with text similar to the user defined query. Their algorithm is dependent on dual deep encoding network. Wu et al. [26] proposed a unsupervised deep video hashing framework used for large scale video retrieval. They integrated video representation with optimal code learning, with provision of an efficient alternative approach to optimize the objective function. Lokoc et al. [27] presented an interactive video retrieval systems using multi-modal search and convenient inspection of results and highlighted query modification statistics. Zhang et al. [28] developed a large scale video retrieval system using a sample image query. They employed convolutional neural networks and Bag of Visual Word for representing video, and a visual weighted inverted index have been introduced by them to improve efficiency and accuracy of the retrieval process. Kordopatis-Zilos et al. [29] introduced a Fine-grained Incident Video Retrieval (FIVR) system aim of which is for a query video, the objective is to retrieve all associated videos. Their method is a single framework containing several retrieval tasks as special case. Rosetto et al. [30] presented a review and results analysis of few retrieval systems. Shen et al. [31] developed a video retrieval system using Similarity-Preserving Deep Temporal Hashing (SPDTH) and their model captures spatio-temporal properties of videos for generation of binary codes. Sauter et al. [32] vitrivr multimedia retrieval stack prepared for participating in 9th Video Browser Showdown (VBS) 2020. In this they presented extra add on facilities to the existing system such as support for classical Boolean queries, metadata filters and added a new object detection module. Zhang et al. [33] proposed a video retrieval system based on query image. They employed keyframe extraction and feature aggregation for retrieval. Sandeep et al. [34] introduced a video retrieval algorithm using hash function generated by tucker decomposition. Thomas et al. [35] proposed a synopsis based video retrieval algorithm using metadata of video such as background, moving object centroid trajectory. Araujo and Girod [36] developed video retrieval algorithm using multi features to retrieve videos from large repositories using query image. Shekar et al. [37] used Local Binary Pattern Variance (LBPV) to retrieve videos. LBPV is a variant of Local binary pattern (LBP), which is LBP augmented with local contrast variation in texture content. Mounika and Khare [38] employed super-resolution and

Histogram of Oriented Gradients (HOG) for content based video retrieval. Their algorithm applies super-resolution to database video frames at interval of 60 to eliminate any degradations in the frame and for the same frames HOG features are extracted and are used for matching with query HOG features and finally to retrieve videos similar to the query.

3 Content Based Video Retrieval (CBVR)

Content based video retrieval algorithm employs keyframe extraction techniques and several feature descriptors for the purpose of retrieval videos similar to the user inputted query. In this section we will discuss several keyframe extraction techniques and the proposed method of keyframe extraction, several feature descriptors available, the proposed method of CBVR and its performance analysis.

3.1 Keyframe Extraction

Video is a rich source of information with lot of redundancy present in it. Processing entire video for any application is complex, tedious job. Instead of processing entire video we can process few frames which are capable of together combinely representing the entire video. Group of such frames is known as keyframes. For developing an efficient keyframe extraction method, it is essential to choose distinguishing features. Statistical features are popular due to their simplicity. Statistical feature vector define statistical distribution of pixels in a frame. Keyframe extraction techniques can be categorized into four types: sequential, clustering, optimization and shot based. Sequential algorithms [39] calculate difference between features of two consecutive frames i.e. current and previous. When the difference is greater than a preset threshold then the current frame is declared as keyframe. The drawback of this algorithm is that the keyframes selected may fail to cover the entire information of video clip efficiently.

In clustering algorithms [40], frames of the video are clustered according to feature similarity using some clustering procedure and the cluster center is chosen as keyframe. Optimization based algorithms [41] select the keyframes by optimizing a desired objective function. Computational complexity and time complexity are high in optimization based algorithms. Shot detection based algorithms [18–20] group similar frames of the video into shots and then extracts keyframes from each shot.

In the present article, we propose a new method of keyframe extraction based on shot boundary detection approach which employ gradient as a statistical feature to calculate dissimilarity and threshold of shot boundary establishment is also designed using statistical parameters mean and standard deviation. The proposed method involves two stages:

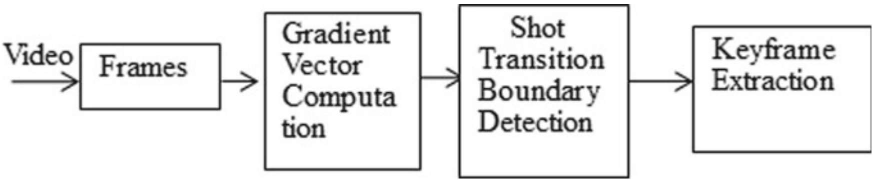


Fig. 2 The proposed method of Keyframe extraction using Gradient

i. In the first stage, shot boundaries are established by calculating gradient.
ii. In the second stage, the last frame of shots identified in the stage (i) are selected as keyframes.

The proposed methodology is attractive due to properties of its high efficiency, less computation and no supervision. The proposed methodology is based on the fact that the difference between the features of consecutive frames varies at the boundary of a shot. Block diagram of the proposed keyframe extraction method is shown in Fig. 2.

Unlike color-based feature, the gradient feature is less susceptible to local illumination changes and camera operations such as- zoom in, zoom out etc. The proposed shot detection method is described below—

Gradient magnitude and direction for each and every frame of the video are extracted. Then, for a video V comprising of N frames let GM_i, GD_i, GM_{i+1}, GD_{i+1} are the gradient magnitude and direction of two consecutive frames F_i and F_{i+1} respectively. Then, the difference in magnitude and direction of ith and $i+1$th frame is given by the following expressions

$$DM(i, i+1) = \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{n} |GM_i(j,k) - GM_{i+1}(j,k)|^2} \tag{1}$$

$$DD(i, i+1) = \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{n} |GD_i(j,k) - GD_{i+1}(j,k)|^2} \tag{2}$$

where, $DM(i, i+1)$ and $DD(i, i+1)$ are dissimilarity between two consecutive frames F_i and F_{i+1} along gradient magnitude and direction respectively.

The dissimilarity in both the magnitude and direction of gradient, as above, is calculated for the entire video sequence. The threshold for establishment of shot transition boundaries is calculated as follows

$$T_M = \mu_M + (\alpha * \sigma_M) \tag{3}$$

$$T_D = \mu_D + (\alpha * \sigma_D) \tag{4}$$

where T_M and T_D are thresholds for dissimilarity matrices *DM* and *DD* respectively.

μ_M, μ_D, σ_M and σ_D are mean and standard deviations of *DM* and *DD* respectively.

α is a constant in the range [0, 1].

Now, the shot transition boundaries are established by comparing the dissimilarities with threshold values obtained using Eqs. (3) and (4). The last frame of each shot is taken as keyframe of that particular shot.

The advantages of the proposed keyframe extraction are due to the incorporation of statistical features, which have proven to be effective feature vectors, it offers ease of computation and approximately robust to noise.

3.2 Feature Vectors for CBVR

Feature descriptors describe visual properties of the video. several broad classifications of features exists such as based on the information employed global and local, Low level and High level. Low level features based on the applications used for, based on the mathematics employed for computation, statistical, Local invariant, textural and transform domain etc. Features such as shape, colour, texture etc. are low level features. High level features are used to define semantic content of the image and they depends on human visual perception. A global feature vectors constructed by taking information present in the entire image into account and global features are used for high-level applications. Features such as invariant moments, shape matrices etc. are examples of Global descriptors. Local feature descriptors are constructed from inform information present in localized portions of the image and local features are used for low-level applications. Features such as SIFT [42], SURF [43], LBP [44], BRISK [20] etc. are examples of local features. Local features offers several benefits such as—

(i) Invariance to scale, rotation, translation, illumination and occlusion.
(ii) They are distinct in nature.
(iii) Even though they captures rich information they are highly compact and efficient.

No matter, whether it is local or global if features are extracted from statistical distribution of pixels present in the frame then they can be said as statistical features. The principal advantages of statistical features are their ease of computation and approximately robust to noise. Several statistical features are available in the literature and some of them are- gradient [45], Color Moments [18], Pearson Correlation Coefficient [18], Histogram of Oriented Gradients (HOG) [38], Laplacian of Gaussian (LOG) [46] etc. Local invariant features are the features that are invariant to image rotation, scale and robust across a substantial range of affine distortion and change in illumination. Many local invariant Features such as Scale Invariant Feature Transform (SIFT) [42], Speeded Up Robust Feature (SURF) [43], Features from accelerated segment test (FAST) [47], Binary Robust Independent Elementary Features (BRIEF) [48] and Binary Robust Invariant Scalable Keypoint (BRISK) [20] etc. are developed in the past. Texture feature descriptor describes spatial arrangement of pixel intensities or colors. The texture features does not depend on the object's size, shape, orientation, and brightness. Textures may employ local information or global information. Examples of texture features that employ local information are Local Binary Pattern [44], Uniform Local Binary Pattern [19], Weber Local Binary Pattern (WLBP) [49], LBP Variance [37] etc. Examples of texture features that employ global information are Log Gabor filters [50], GLOGTH [51] etc.

The efficiency of CBVR system depends on efficient feature vectors selection and extraction. Wide number of feature descriptors are available in an image. Mostly low level features are extracted. In following we will discuss few feature vectors used in the proposed method of content based video retrieval (CBVR).

3.2.1 Laplacian of Gaussian (LOG)

To find rapid change areas in a frame Laplacian filters have been used. Noise sensitivity is more in any derivative filter. Since, Laplacian is a derivative filter, to reduce noise sensitivity, a Gaussian smoothing will be applied to the frame and then on smoothed version of frame Laplacian filtering will be applied. The process combinely is known as Laplacian of Gaussian (LOG). The LOG filter is used to highlight edges in a frame. The LOG filter detects or highlights edges in a frame through computation of second order spatial derivative. LOG yields a zero response for an image with constant intensity. The filter response will be—

(i) Zero at spatial locations far away from the edge.
(ii) Positive at darker portions located side to the edge.
(iii) Negative at brighter portions located side to the edge.

3.2.2 Histogram of Oriented Gradients (HOG)

Dalal et al. [15] introduced Histogram of Oriented Gradients (HOG). HOG descriptor gives us the number of times a particular gradient orientation appears in a localized portion of an image. The theme behind HOG is appearance and shape of a local object can be better described by the distribution of intensity gradients. The attractive feature of HOG is that it is invariant to geometric and photometric transformation. HOG efficiently captures the shape of object and foreground, structure of edge.

3.3 The Proposed Method

The block diagram of the proposed content based video retrieval algorithm is shown in Fig. 3.

In offline processing, the videos stored in database are accessed one by one. Each video of database undergo shot detection and keyframe extraction. Now the

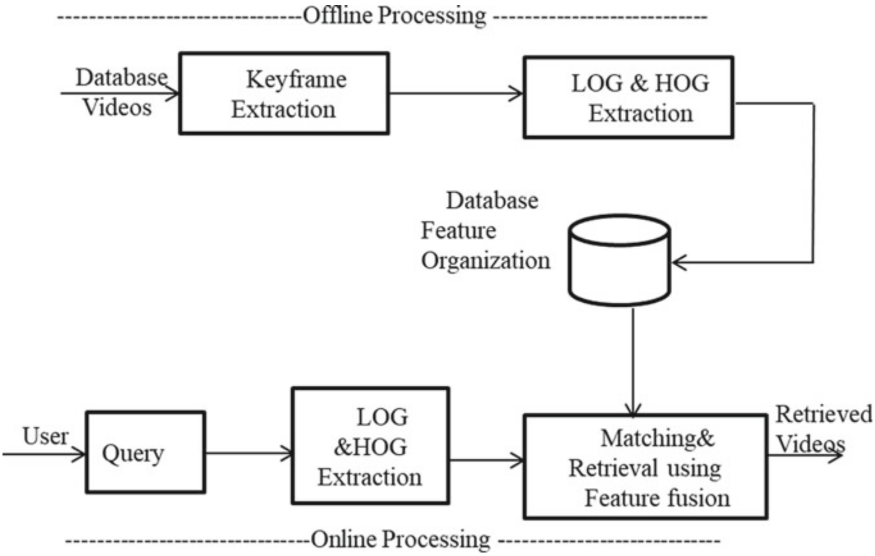


Fig. 3 The Proposed content based video retrieval system

features Laplacian of Gaussian (LOG) and Histogram of Oriented Gradients (HOG) of keyframes are extracted. The LOG and HOG features of database video keyframes together are used to represent the database videos. In online processing, user inputs query which is a sample image along with the text describing the category of the video. Then, LOG and HOG features of query image are extracted, used to represent the query image and these features are searched to find a matched features in the database. The database videos whose LOG and HOG features of keyframes matched with query frame's LOG and HOG features are retrieved separately to get final retrieval result.

The LOG and HOG features of the query frame are matched with the features of database keyframes by performing matching. For the purpose of matching we employed euclidean distance in the proposed retrieval algorithm. Let K_d= [K_1, $K_2 \ldots K_N$] denote the keyframes of a database video V_d. Then euclidean distance is calculated between the features of query frame Q_i and features of each K_i, $i \epsilon$ [*1*, *N*] of database video V_d to perform matching and its mathematical equation is given as—

$$ED = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (F_q(i,j) - F_K(i,j))^2} \tag{5}$$

where,

ED Euclidean distance

$F_q(i, j)$ LOG/HOG feature of query frame q at spatial location (i, j)
$F_K(i, j)$ LOG/HOG result of keyframe K at spatial location (i, j).

Euclidean distance between both the LOG and HOG features of query frame and database video keyframes are calculated separately. Then corresponding database videos for which the euclidean distance of either LOG/HOG is less considered as matched one.

3.4 Performance Analysis

The proposed method have been experimented on the entire dataset developed by Mounika and Khare [38] which is publicly available at https://sites.google.com/site/mounikabrv3/research-profile. Quality of any proposed algorithm will be judged by analyzing performance of the method. Performance analysis can be done either qualitatively or quantitatively. Qualitative performance analysis is a subjective way of measuring quality and it depends on user and varies from user to user. Quantitative performance analysis is an objective way of measuring quality and it does not depend on user and so leads to impartial judgment. Performance of the proposed retrieval algorithm is evaluated both qualitatively and quantitatively and their results are compared with other state-of-art methods [34–38].

3.4.1 Qualitative Performance Analysis

In this section for an example query frame of dataset, top ten retrieved results extracted by the proposed method when only HOG feature used and only LOG feature used and when both the features used have been presented in Figs. 4, 5, 6 and

Fig. 4 An example query frame of serials category video of dataset

Fig. 5 Representative frames of top ten retrieved video results obtained by the proposed method when only HOG employed for query in Fig. 4

Fig. 6 Representative frames of top ten retrieved video results obtained by the proposed method when only LOG employed for query in Fig. 4

7 respectively. For the same query, top ten retrieved results extracted by the other methods the other methods [34–38] have been presented. The proposed retrieval method has been compared with five different state-of-art methods namely-Sandeep

Fig. 7 Representative frames of top ten retrieved video results obtained by the proposed method using LOG and HOG feature fusion for query in Fig. 4

et al. [34], Thomas et al. [35], Araujo and Girod [36], Shekar et al. [37] and Mounika and Khare [38]. The above methods are well known, recent feature based methods of this field and all the methods employed different features, which made us to choose these methods for comparative analysis of the proposed method. The top ten retrieval results obtained by the methods- Sandeep et al. [34], Thomas et al. [35], Araujo and Girod [36] and Shekar et al. [37] and Mounika and Khare [38] have been presented in Figs. 9, 10, 11, 12 and 13 respectively.

From Figs. 5, 6 and 7, we can observe that when only HOG feature is used only 4 videos similar to query are retrieved, when only LOG feature is used only 6 videos similar to the query are retrieved and when both the features used then 8 videos similar to the query are retrieved as result. The performance is improved with fusing two features namely HOG and LOG. From Figs. 8, 9, 10, 11 and 12 we can clearly observe that the method by Sandeep et al. [34] retrieved 4 videos, the method by Thomas et al. [35] retrieved 3 videos, the method by Araujo and Girod [36] retrieved 5 videos, the method by Shekar et al. [37] retrieved 5 videos, the method by Mounika and Khare [38] which is base of the proposed retrieval method retrieved 5 videos relevant to query either in sense of object or in sense of background or even both with slight changes of illumination.

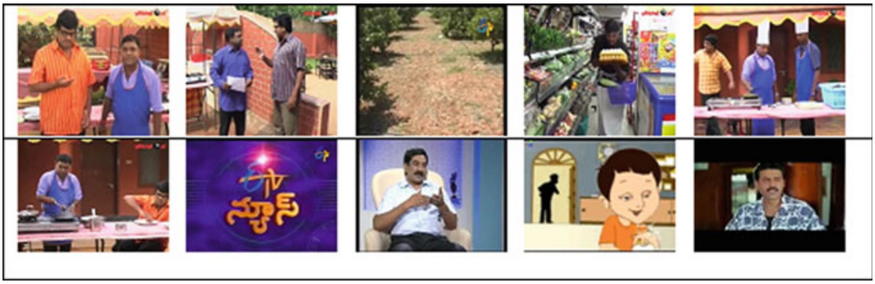


Fig. 8 Representative frames of top ten Retrieved video results obtained by Sandeep et al. [34] method for query in Fig. 4

Fig. 9 Representative frames of top ten Retrieved video results obtained by Thomas et al. [35] method for query in Fig. 4

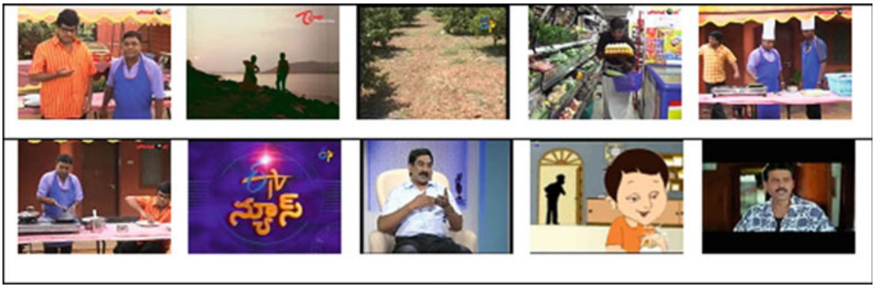


Fig. 10 Representative frames of top ten Retrieved video results obtained by Araujo and Girod [36] method for query in Fig. 4

Fig. 11 Representative frames of top ten Retrieved video results obtained by Shekar et al. [37] method for query in Fig. 4

Fig. 12 Representative frames of top ten Retrieved video results obtained by Mounika and Khare [38] method for query in Fig. 4

On an overall the proposed method given good qualitative results. The benefit comes from two reasons one is fusing multiple features. In any case, if one feature fails another feature may success and the other is addition of one more field in query describing the category of the video searched for. The representative frame of the retrieved video clip in all Figs. 5, 6, 7, 8, 9, 10, 11 and 12.

3.4.2 Quantitative Performance Analysis

To analyse performance of the proposed method we chosen five different quantitative performance measures. All the chosen five performance measures are designed based on reference of ground truth. For our work, we have built ground truth manually. The proposed method performance is evaluated with five parameters they are- Precision, Recall, Jaccard index, Accuracy and Specificity. They are mathematically given as—

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$J = \frac{TP}{TP + FP + FN} \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

where,

True Positive (TP) Number of videos that are identified as relevant to query by both the ground truth and the algorithm.
True Negative (TN) Number of videos that are identified as irrelevant to query by both the method and ground truth.
False Positive (FP) Number of videos that are incorrectly retrieved as relevant by the method but not present in the ground truth.
False Negative (FN) Number of videos that are relevant to query as per the ground truth but not present in the retrieved result of algorithm

From results furnished in Table 1 and Fig. 13 we can clearly observe that the proposed CBVR method with incorporation of feature fusion with HOG and LOG is

Table 1 Performance comparison of the proposed CBVR method with single feature and with feature fusion with other state-o-art CBVR methods [34–38]

Performance parameter	Only HOG	Only LOG	The proposed CBVR method
Precision	32.01	25.4	59.91
Recall	32.03	25.20	59.65
Accuracy	25.65	20.67	58.56
Jaccard index	23.35	23	82.79
Specificity	22.12	15.8	37.97

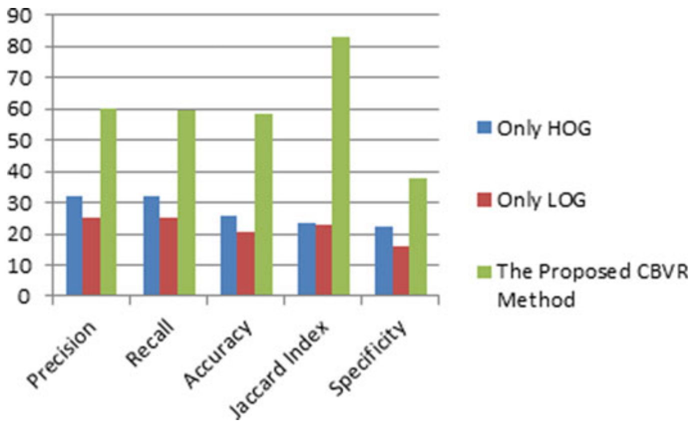


Fig. 13 Average Value of Performance Measures Versus the feature set for the proposed CBVR method

performing far better compared to incorporation of a single feature either only HOG or LOG in terms of Precision, Recall, Accuracy, Jaccard Index and Specificity.

From results furnished in Table 2, Fig. 14, we can clearly observe that the proposed CBVR method given better performance than the other methods [34–38] in terms of Precision, Recall, Accuracy, Jaccard Index and Specificity.

Table 2 Performance comparison of the proposed CBVR method with other state-of-art CBVR Methods [34–38]

Method	Precision (%)	Recall (%)	Accuracy (%)	Jaccard index (%)	Specificity (%)
Sandeep et al. [34] method	16.73	23.33	15.17	11.00	12.50
Thomas et al. [35] method	4.42	4.42	7.62	2.42	11.04
Araujo and Girod [36] method	20.83	20.83	22.22	14.61	36.10
Shekar et al. [37] method	32.65	37.07	26.91	23.68	25.44
Mounika and Khare [38] method	37.20	37.20	29.88	25.74	25.21
The proposed CBVR method	59.91	59.65	58.56	82.79	37.97

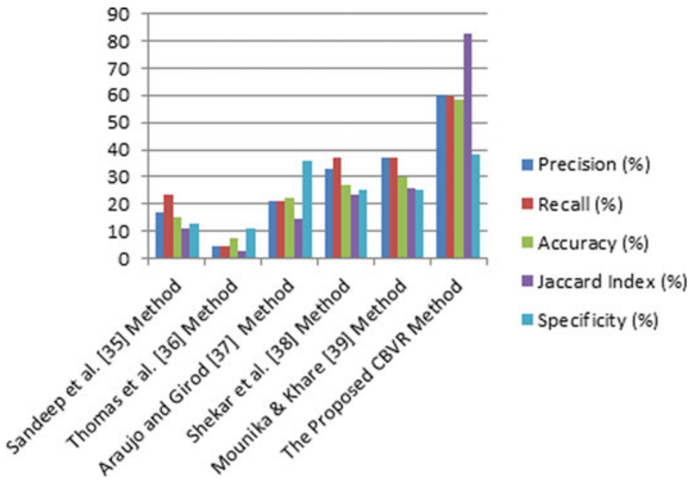


Fig. 14 Average Value of Performance Measures Versus the state-of-art methods [34–38] and the proposed CBVR method

4 Applications

Content based video retrieval founds its applications widely in several computer vision tasks such as innovative smart city applications such as surveillance, traffic monitoring, crowd monitoring, activity recognition etc. educational applications such as e-learning, bio medical applications such as monitoring different health conditions MRI Scan, surgeries to improve health care services and entertainment application such as Video-on demand etc.

5 Conclusion

This chapter discussed concept of feature fusion for content based video retrieval. Feature fusion is beneficial since, if one feature fails in capturing query content and matching it with database, another feature may success. The effectiveness of feature fusion was explained with the help of the proposed Content based video retrieval (CBVR) method. The proposed method two different features namely HOG and LOG. The proposed method's performance improved more compared to the proposed method's performance with a single feature. The application of feature fusion is not only limited to CBVR domain only. Recent advancements lead to the application of fusion concepts in several different computer vision fields. Since, different categories of videos own different properties. A single feature cannot hold the characteristics of those different videos. Hence, feature fusion finds well applicable for different video processing fields especially to CBVR and the CBVR domain founds its application in wide range of computer vision fields. Hence, the CBVR combined with feature fusion proven to be an effective research in the domain of video processing.

References

1. Mashtalir, S., Mashtalir, V.: Spatio-temporal video segmentation. In: Advances in Spatio-Temporal Segmentation of Visual, pp. 161–210. Springer, Cham (2020)
2. Xu, K., Wen, L., Li, G., Bo, L., Huang, Q.: Spatiotemporal cnn for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1379–1388 (2019)
3. Tian, Y., Cheng, G., Gelernter, J., Yu, S., Song, C., Yang, B.: Joint temporal context exploitation and active learning for video segmentation. Pattern Recogn. **1**(100), (2020)
4. Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 440–448. Springer, Cham (2019)
5. Zhong, S.H., Wu, J., Jiang, J.: Video summarization via spatio-temporal deep architecture. Neurocomputing **7**(332), 224–235 (2019)
6. Zhang, K., Wang, L., Liu, D., Liu, B., Liu, Q., Li, Z.: Dual Temporal Memory Network for Efficient Video Object Segmentation (2003). arXiv preprint arXiv:2003.06125
7. Ahmadi, N., Akbarizadeh, G.: Iris tissue recognition based on GLDM feature extraction and hybrid MLPNN-ICA classifier. Neural Comput. Appl. **32**(7), 2267–2281 (2020)
8. Wu, Y.: Research on feature point extraction and matching machine learning method based on light field imaging. Neural Comput. Appl. **31**(12), 8157–8169 (2019)
9. Qin, Y., Zou, J., Tang, B., Wang, Y., Chen, H.: Transient feature extraction by the improved orthogonal matching pursuit and K-SVD algorithm with adaptive transient dictionary. IEEE Trans. Industr. Inf. **16**(1), 215–227 (2019)
10. Wang, R., Shi, Y., Cao, W.: GA-SURF: a new speeded-up robust feature extraction algorithm for multispectral images based on geometric algebra. Pattern Recogn. Lett. **1**(127), 11–17 (2019)
11. Janwe, N., Bhoyar, K.: Semantic concept based video retrieval using convolutional neural network. SN Appl. Sci. **2**(1), 80 (2020)
12. Francis, D., Anh Nguyen P, Huet B, Ngo CW. Fusion of multimodal embeddings for ad-hoc video search. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0–0 (2019)
13. Zhang, J., Mei, K., Zheng, Y., Fan, J.: Exploiting mid-level semantics for large-scale complex video classification. IEEE Trans. Multimed. **21**(10), 2518–2530 (2019)
14. Prathiba, T., Kumari, R.S.: Content based video retrieval system based on multimodal feature grouping by KFCM clustering algorithm to promote human–computer interaction. J. Ambient Intell. Humaniz. Comput. **13**, 1–5 (2020)
15. Zeng, J., Liu, M., Fu, X., Gu, R., Leng, L.: Curvature bag of words model for shape recognition. IEEE Access. **29**(7), 57163–57171 (2019)
16. Agosti, M., Marchesin, S., Silvello, G., Vezzani, F., Di Nunzio, G.M., Tellez, D., Hoppener, D., Verhoef, C., Grunhagen, D., Nierop, P., Drozdzal, M.: Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval. ACM Trans. Inf. Syst. **1**(1) (2020)
17. Song, G., Tan, X.: Deep code operation network for multi-label image retrieval. Comput. Vis. Image Underst. **1**(193), (2020)
18. Bommisetty, R.M., Prakash, O., Khare, A.: Keyframe extraction using Pearson correlation coefficient and color moments. Multimed. Syst. **18**, 1–33 (2019)

19. Mounika, B.R., Prakash, O., Khare, A.: Key frame extraction using uniform local binary pattern. In: 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), pp. 87–91. IEEE (2018)
20. Khare, A., Mounika, B.R., Khare, M.: Keyframe extraction using binary robust invariant scalable keypoint features. In: Twelfth International Conference on Machine Vision (ICMV 2019), vol. 11433, p. 1143308. International Society for Optics and Photonics (2020)
21. Mounika, B.R., Khare, A.: Shot boundary detection using second order statistics of gray level co-occurrence matrix. Res. J. Comput. Inf. Technol. Sci. **5**, 1–7 (2017)
22. Abed, R., Bahroun, S., Zagrouba, E.: KeyFrame extraction based on face quality measurement and convolutional neural network for efficient face recognition in videos. Multimedia Tools Appl. **6**, 1–22 (2020)
23. Li, Y., Kanemura, A., Asoh, H., Miyanishi, T., Kawanabe, M.: Multi-Sensor integration for key-frame extraction from first-person videos. IEEE Access. **9**(8), 122281–122291 (2020)
24. Lokoč, J., Bailer, W., Schoeffmann, K., Münzer, B., Awad, G.: On influential trends in interactive video retrieval: video browser showdown 2015–2017. IEEE Trans. Multimedia **20**(12), 3361–3376 (2018)
25. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9346–9355 (2019)
26. Wu, G., Han, J., Guo, Y., Liu, L., Ding, G., Ni, Q., Shao, L.: Unsupervised deep video hashing via balanced code for large-scale video retrieval. IEEE Trans. Image Process. **28**(4), 1993–2007 (2018)
27. Lokoč, J., Kovalčík, G., Souček, T., Moravec, J., Čech, P.: VIRET: A video retrieval tool for interactive known-item search. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval, pp. 177–181 (2019)
28. Zhang, C., Lin, Y., Zhu, L., Liu, A., Zhang, Z., Huang, F.: CNN-VWII: an efficient approach for large-scale video retrieval by image queries. Pattern Recogn. Lett. **15**(123), 82–88 (2019)
29. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, I.: FIVR: Fine-grained incident video retrieval. IEEE Trans. Multimedia **21**(10), 2638–2652 (2019)
30. Rossetto, L., Gasser R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., Vrochidis, S.: Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. IEEE Trans. Multimedia. (2020)
31. Shen, L., Hong, R., Zhang, H., Tian, X., Wang, M.: Video retrieval with similarity-preserving deep temporal hashing. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **15**(4), 1–6
32. Sauter, L., Parian, M.A., Gasser, R., Heller, S., Rossetto, L., Schuldt, H.: Combining boolean and multimedia retrieval in vitrivr for large-scale video search. In: International Conference on Multimedia Modeling, pp. 760–765. Springer, Cham (2020)
33. Zhang, C., Hu, B., Suo, Y., Zou, Z., Ji, Y.: Large-scale video retrieval via deep local convolutional features. Adv. Multimedia. **9**, 2020 (2020)
34. Sandeep, R., Sharma, S., Thakur, M., Bora, P.K.: Perceptual video hashing based on Tucker decomposition with application to indexing and retrieval of near-identical videos. Multimedia Appl. **75**(13), 7779–7797 (2016)
35. Thomas, S.S., Gupta, S., Venkatesh, K.S.: Perceptual synoptic view-based video retrieval using metadata. SIViP **11**(3), 549–555 (2017)
36. Araujo, A., Girod, B.: Large-scale video retrieval using image queries. IEEE Trans. Circuits Syst. Video Technol. **28**(6), 1406–1420 (2018)
37. Shekar, B.H., Uma, K.P., Holla, K.R.: Video clip retrieval based on LBP variance. Procedia Comput. Sci. **1**(89), 828–835 (2016)
38. Mounika, B.R., Khare, A.: Content based video retrieval using histogram of gradients and frame fusion. In: Twelfth International Conference on Machine Vision (ICMV 2019), vol. 11433, p. 114332J. International Society for Optics and Photonics (2020)
39. Shi, Y., Yang, H., Gong, M., Liu, X., Xia, Y.: A fast and robust key frame extraction method for video copyright protection. J. Electr. Comput. Eng. (2017)

40. Kannappan, S., Liu, Y., Tiddeman, B.: DFP-ALC: automatic video summarization using distinct frame patch index and appearance based linear clustering. Pattern Recogn. Lett. **120**, 8–16 (2019)
41. Liu, X.M., Hao, A.M., Zhao, D.: Optimization-based key frame extraction for motion capture animation. Vis. Comput. **29**(1), 85–95 (2013)
42. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
43. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: speeded up robust features. In: European Conference on Computer Vision, pp. 404–417. Springer, Berlin, Heidelberg (2006)
44. Porebski, A., Vandenbroucke, N., Macaire, L.: Haralick feature extraction from LBP images for color texture classification. In: 2008 First Workshops on Image Processing Theory, Tools and Applications, pp. 1–8. IEEE (2008)
45. Liu, C.L.: Normalization-cooperated gradient feature extraction for handwritten character recognition. IEEE Trans. Pattern Anal. Mach. Intell. **29**(8), 1465–1469 (2007)
46. Yaji, G.S., Sarkar, S., Manikantan, K., Ramachandran, S.: DWT feature extraction based face recognition using intensity mapped unsharp masking and laplacian of gaussian filtering with scalar multiplier. Procedia Technol. **1**(6), 475–484 (2012)
47. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision, pp. 430–443. Springer, Berlin, Heidelberg (2016)
48. Calonder, M., Lepetit, V., Strecha, C., Brief, F.P.: Binary robust independent elementary features. In: European Conference on Computer Vision, pp. 778–792 (2010)
49. Liu, F., Tang, Z., Tang, J.: WLBP: Weber local binary pattern for local image description. Neurocomputing **23**(120), 325–335 (2013)
50. Wang, W., Li, J., Huang, F., Feng, H.: Design and implementation of Log-Gabor filter in fingerprint image enhancement. Pattern Recogn. Lett. **29**(3), 301–308 (2008)
51. Poongothai, E., Suruliandi, A.: Global and local oriented gabor texture histogram for person re-identification. Braz. Arch. Biol. Technol. **62** (2019)

Building the World of Internet of Things

Seema Sahai, Richa Goel, and Gurinder Singh

Abstract When technology started growing, there was a vision of a world, by many leading technologists, that objects of all kinds would be behaving and interacting like humans. They would be able to analyse, react, sense and communicate with each other as well as humans in exactly the same way as humans do. All this done by gathering data. Any action reveals a set of data about it and using that data one can predict and modify future acts. This is a very generic way of explaining the phenomenon. The process is however, quite complicated. There are a whole set of technologies which allow the gathering of data and creating information. The "thing" in the internet of things may have several instances, such as a human being with a pace maker, a wildlife animal with a biochip transponder, a built-in vehicle or some other natural or man-made object that can be given an Internet Protocol (IP) address that can transfer data over a network. IOT has an impact on various aspects of businesses like analytics, security, customer information etc. Today, most businesses are trying to integrate IoT enabled capabilities into their strategies. But before that a general understanding of the technologies is very important. Communication in Organisations is done through networks which can be said to be done through electronic signals. An example of this would be the wireless communication mechanism that exists and allows for large dataflow. Sensors are also devices which "sense" certain conditions like in cars, indicators are there for fuel level, battery, temperature as well as closeness of an object. These are one example. Sensors are now seen in many devices like air conditioners, heaters, lighting etc. Analytical instruments are used by organisations to enhance their capacity to describe, predict and take advantage of certain connections or links. These methods are said to be using augmented technology. When procedures are taken in a prescribed way of action with the intervention of humans using technology and methods according to prescribed

S. Sahai (✉) · R. Goel · G. Singh
Amity International Business School, Amity University Uttar Pradesh, Noida, India
e-mail: ssahai@amity.edu

R. Goel
e-mail: rgoel@amity.edu

G. Singh
e-mail: gsingh@amity.edu

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_5

action, it is known as augmented behaviour. Machine-to-machine interfaces delete human faults in a reliable manner. There are also certain commonly prescribed norms for action known as standards. These standards enable compatibility with different platform using IOT. This chapter will give an insight to all the underlying technologies of IOT. Starting from Networks to sensors, standards, Augmented Intelligence, Augmented behaviour.

Keywords Internet of Things · Technology · Data · Network · Communication

1 Introduction: The Information Revolution

Whenever a drastic change occurs, we call it a revolution. The Information Revolution has been taking place for the past few years and is at its peak at this time. The root of this change is Computer technology. This change has been touching the lives of almost everyone in the world. Workers are being replaced by intelligent bots and robots. There is a fear that people will lose their jobs to this intelligent devices which have been built by humans themselves. Here we need to understand that with benefits of the technological revolution, there are disadvantages of it too, which has a lot of psychological and social effects on people.

Bojanova [2] has been said that the entire business processes around the world has been transformed by current technological innovations. It has had its effects on production, manufacturing and services as well. The major technologies that are driving this change are 3D printing, autonomous vehicles and various computer interfaces. This has led organisations to acquire new systems and employees have had to upgrade their skill to remain employed.

According to Wyman [5], digitisation is very difficult to measure. He is of the opinion that domains of life be it economic, social or cultural will be affected by digital services. Network will be accessed by all age groups, all classes and all countries and this will have a tremendous effect on the total lifestyle.

In the past decade smart technology, including social networking and big data have given ways to organizations to expand their horizons [3]. This is further enhanced by greater mobility through Internet of Things and 5G mobile technology, which is opening new markets. IOT has given rise to innovative technologies called 'smart 'technologies and therefore a smart revolution is changing everyone's lives.

Mühleisen [4] says that with the increase and development of computing power many more people are participating in the digital economy. This needs that policies are also devised in such a way that the benefits of digital revolution are exploited with minimal job dislocation. He also mentioned that adaption of society to a disruptive technology is much faster than the adoption of the technology. The main reason being the long time it takes for the output growth to accelerate. The benefits of such revolutions are actually perceived much later.

The digital revolution is marked by cloud computing being a fundamental feature [6]. The organisations could simply remove human intelligence from work tasks and displace the work. They could also, instead of displacing, augment the workforce with intelligent tools. It is possible to choose the kind of society and organisation wants.

Baker [1] opines that revolutions are bound to create more jobs than it is going to destroy. However, he feels that this digital or information revolution will not be the same. To avoid this the education system also has to undergo a drastic change. It has to move on to and active way of education than the age old passive way. Personal and collaborative skills have become more important and they should be focussed on. He feels that machines will take over routine tasks for sure and non-routine tasks also to some extent and this is where the job market will suffer (Fig. 1).

The World Economic Forum has predicted that by the end of 2025, the number of subscription for %G will be covering about 65% of the world's population. It is

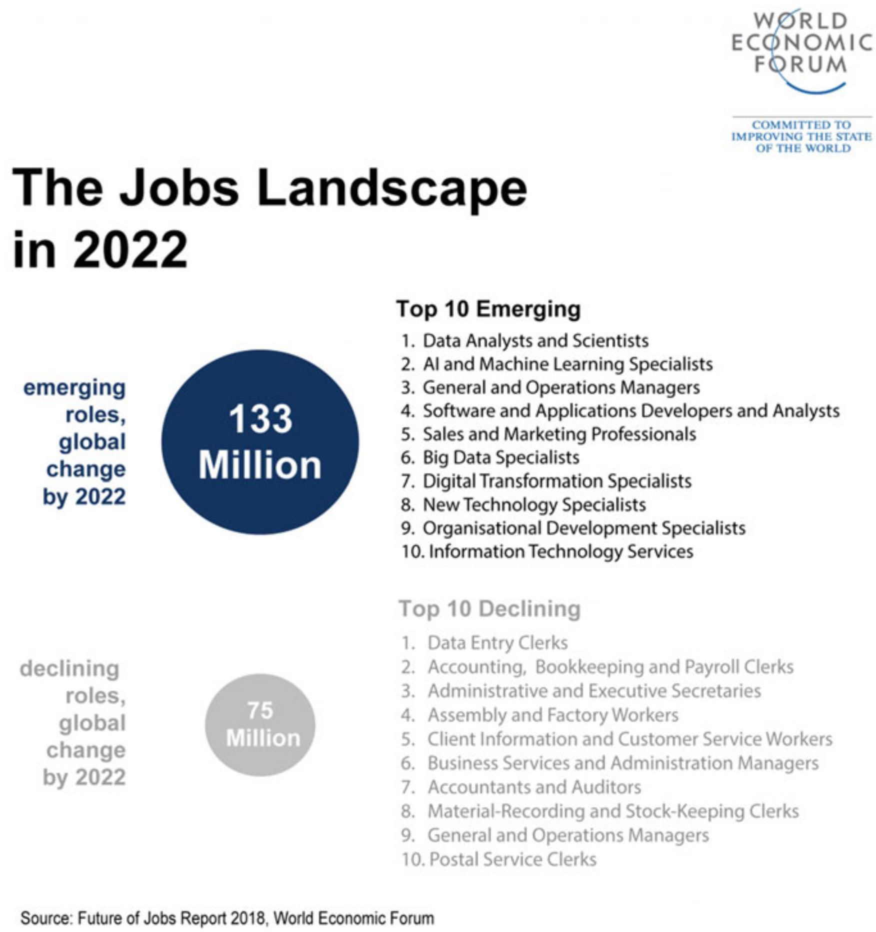


Fig. 1 World Economic Forum, Future of jobs Report 2018

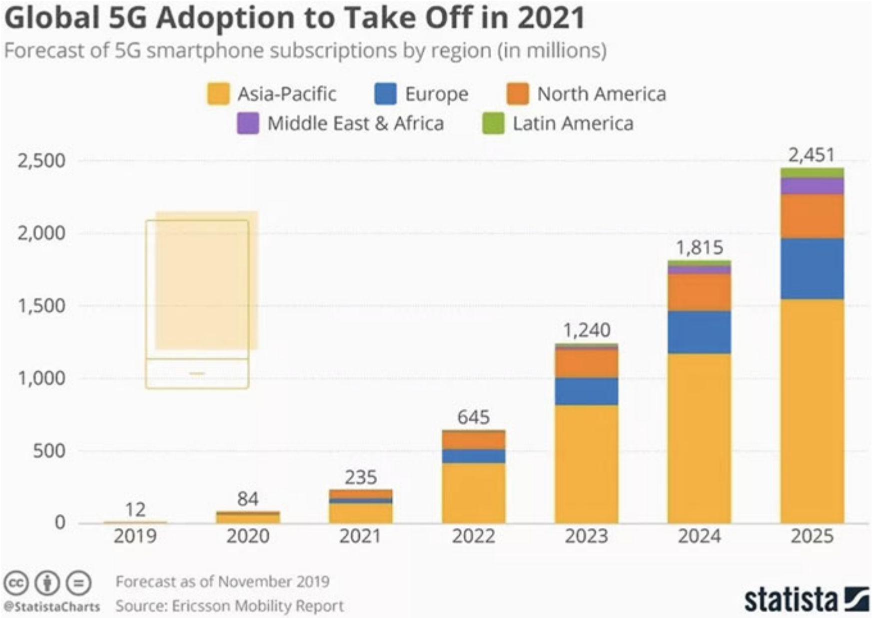


Fig. 2 Forecast of Global 5G adoption by Statistica

expected the there will be 6 billion cellular IOT connections. Businesses will be able to surge forward by adopting new technologies and IOT itself will generate great economic value.

IoT is a case of strategic use for the 5G network. This is because traditional mobile networks tend to struggle to keep up with the booming market's capital demands for IoT applications.

The three major 5G networks use IoT support scenarios as specified by the International Telecommunications Union (ITU) are enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable, low-latency communications (URLLC) (Fig. 2).

2 Sensors

Sensors have a technology, which allows machines to see, recognize and comprehend the world without human intervention. Sensors when attached to an inanimate thing make it behave like humans. They basically take the actions of humans and create information out of it. It of course does not do it in isolation; it takes the help of modem chips, power, microcontrollers and many other devices. This information is sent to an electrical circuit in the form of an electrical signal.

Fig. 3 Functionality of sensors. *Source* Shuttershock

A sensor then requires an actuator, which converts the electrical signals into actions. It is actually a conversion of one form of energy into another. Energy of different forms gets converted to electrical energy and the into another form. There are many types of sensors. Each sensing different types of energy. All the sensors combined together create a complex loop using different kinds of information. A good example would be a home security system, which checks temperature, humidity, light etc. These are all different kinds of sensors combined with sensors that detect motion and behavior. These sensors improve security of a home (Fig. 3).

Sensors can be said to be active and passive. Passive sensors are sensors, which receive energy from an external source. Sun is an example, which provides external energy for remote sensing. The remote sensing devices are all passive sensors. Sun's energy may not be available at night but other sources of energy are and this is how these devices work at night.

Active sensors on the other hand, provide their own energy, which is directed towards the target under observation. An example of active sensors is the RADAR system (Fig. 4).

The choice of the sensor depends on the function that is to be performed. Some of the determinants could be noise, resolution, accuracy, range, repeatability and selectivity amongst others.

The reasons why sensors are being adopted in IOT are many. Some of them are that the cost of sensors is constantly decreasing and in turn these sensors are becoming more intelligent and smart. The computational power of sensors is said to double every three years. It is to be remembered that sensors do not work in isolation but

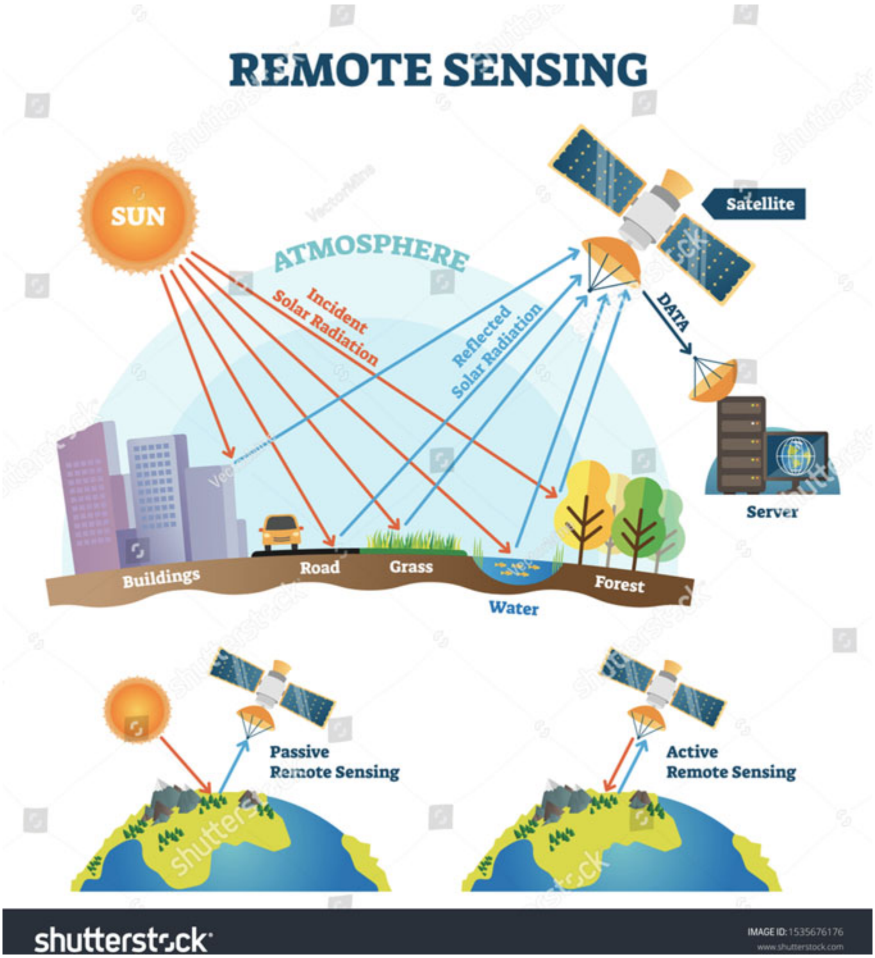


Fig. 4 Active and passive sensing. *Source* Shuttershock

with many other interconnected devices. The size of the sensors are also constantly decreasing, making it a very compact chip, which can be, installed anywhere.

In spite of the above benefits, there are a number of challenges in implementation. The first challenge is that of security. The low processing power and low memory makes it difficult to ensure security in the sensors. Cryptographic algorithm is one way to enable some security in the sensors. Second challenge is the amount of power that is consumed by these sensors. Power sources have been proven to be constant but expensive. Through innovation some sensors can run on batteries for long times like 10 years. This is one way to overcome the challenge. Looking for alternative sources of power like solar power is also an option but it is expensive. Another challenge is that of adaptability to various applications. Many sensors are yet to be innovated to be compatible.

Sensor is an essential part and plays a major role in the functioning of many companies of today. They alert the concerned people about possible issues beforehand and if the issues get elevated, they help in businesses to conduct predictive maintenance and prevent costly downtime. It is also possible to analyze sensor data for patterns that allow the top level to gain understanding into key trends and make informed proof-based decisions. Some popular uses of sensors are as follows:

(i) Temperature sensors. They measure a system's amount of heat energy so that they can perceive temperature changes and translate those changes into data. Machinery utilized in fabrication often necessitate specific ambient and device temperature levels. Similarly, ground temperature is a critical factor for agricultural crop production.
(ii) Humidity sensor. These sensor determine the amount of water vapor along with other gases present in the atmosphere. Humidity sensors are typically used in industrial and residential heating, windows, and air-conditioning (HVAC) systems. These can be used for monitoring and forecasting weather in many other fields including hospitals, and meteorological stations.
(iii) A pressure sensor detects changes in liquids and gases. The sensor senses those changes as the pressure increases, and communicates them to linked devices. Popular use cases provide surveillance of the leakage that can result from degeneration. Pressure sensors are also suitable in the expansion of water systems, since it is simple to detect variations or falls in the pressure.
(iv) Proximity sensors are used to detect objects near to the sensor, without touch. Such types of sensors often discharge electromagnetic fields, or radiations. In the retail field, a proximity sensor will detect the movement between a customer and a produce he or she is interested in. They can also be seen on the manufacturing lines of chemical-, food- and many other forms of industry.
(v) Level Sensors. For measuring the substance levels such as liquids, powders, granular materials and pressure sensors are used. Most companies use level sensors like oil processing, water treatment, and soda factories, as well as food production. Waste management systems have a common use case, as level sensors can measure the amount of waste in a garbage can or dumpster.
(vi) The accelerometer calculates an object's acceleration, i.e. the rate of change in the body's velocity in relation to the time. Accelerometers, too, can detect changes in gravity. Use cases for accelerometers include smart pedometers and monitoring of the moving vehicles. These can also be used as protection against theft to warn the system when moving an item that is supposed to be stationary.
(vii) Gyroscope sensors measure angular velocity or rate, also defined as velocity and gyration measurements around an axis. Use cases include vehicle technologies. Unique use cases comprise movement tracking in video games, and camera shake detection.

(viii) Gas Sensors. Such types of sensors track and identify alterations in air quality like poisonous, inflammable, or hazardous gasses. Industries using gas sensors include mining, oil and gas, chemical testing, and manufacturing. The famous carbon dioxide detectors that are found in many homes are a typical case of consumer use.

(ix) Infrared sensors. Such types of sensors sense atmospheric features by also releasing or detecting radiation from the infrared. They can also calculate the heat, which objects are emitting. Infrared sensors are used in a variety of IoT applications such as healthcare, as they promote blood flow and control blood pressure. Televisions use infrared sensors to view distant signals.

(x) The optical sensors translate light rays to electrical signals. Optical sensors have multiple applications, and use cases. Throughout the automobile industry, cars are using optical sensors to detect signs, obstacles and other objects that a driver would encounter while driving or parking. Optical sensors play a major role in the production of driverless cars. On smart phones, the optical sensors are very common. The ambient light sensors, for example, will extend the battery life. The biomedical sector also uses optical sensors including breath analysis and heart rate monitoring.

3 Networks

It has been understood that sensors are an important part of IOT but it needs to be understood that the data that sensors create needs to be communicated to devices or human and for this a network is required. It is therefore necessary that for transmission of data a network is required. In a network many networking devices are attached. These devices could be routers, modem, hubs, gateways, switches etc.

The transmission of data is done through the network by uniquely identifying the devices or things in it. For this a unique name is given to each device and a network protocol is used to identify the devices. Protocols can be proprietary or open. Proprietary protocols are those which are used by specific hardware and software, whereas, open protocols are more scalable and can easily be used across heterogeneous devices.

One such open protocol is the Internet Protocol (IP). IP gives a unique address to each device connected to the Internet. Since the estimated number of devices connected to the Internet is assumed to be 50 billion or more, the IPv6 has been a key enabler of IOT.

Networks can be wired or wireless. Since the world of Internet most devices are mobile so wireless network is a preferred network. They are convenient and have continuous connectivity. However, when it comes to reliability, wired networks are better. People chose the type of network depending on the range. When the range is small a wireless network performs better. But when network is slightly large the wired network is preferred due to its reliability. This is why a LAN is usually a wired network. A number of LAN networks further create a WAN.

The rate in which data has to be transferred is also one reason for selecting a particular type of network. 5G and 4G are used in IOT because they have high rates of data transfer. Bluetooth and Low-power Wi-Fi is used where data is not transferred at a high rate.

Microwave Access (WiMAX) is also a method used for networks. WiMAX2 has a maximum speed of 1 Gbps.

The LTE or Long Term Evolution is used for WAN and its latest version also supports 1 Gbps of data.

With the improvement of network technologies came in many challenges such as security, power consumption, penetration and interconnections. In terms of interconnections all devices are not compatible with each other. The compatibility issue has raised a number of questions on complexity and affordability.

Many of the high bandwidth technologies have limited penetration like LTE. The security of the network is under massive threat due to the number of sensors attached to it. It is very difficult to maintain data integrity and being energy efficient at the same time.

A continuous power source is required by all IOT devices. This at times proves to be a problem. A number of activities combined allows it to remain energy efficient.

With the advent of the IoT, an explosion of data is inevitably being applied to this new world's networks and data collection, storage, and analytics systems.

Connectivity and networks require an IoT framework to fit in with. It's no surprise, for example, that a company such as Cisco claims a leading position in the IoT space when its core business is still very much about networks, even if it is evolving. Although various new types of network and communication technology and, of course, the cloud allowed or increased better. The growth of the Internet of Things (and vice versa, since the cloud and IoT are closely linked), IoT also affects network and infrastructure technology evolution.

As well as demands for increased power and network awareness as a result of digital transformation, there is also a new infrastructure paradigm whereby electricity, computing, networking and cloud technologies shift closer to the devices themselves, closer to the edge of the network. The name is Edge Computing and it's getting bigger. Cisco is also a main force in here.

Within the Internet of Things, there is not a single type of network and computer infrastructure that solves all the data and analytics problems and possibilities that various implementation scenarios pose us.

Also, there is no single type of networking solution that links individuals, stuff and networks in shorter ranges. As you know, IoT is a paragliding concept for many possible applications in many possible industries where several technologies may be more important than others and de facto it is always an IoT combination.

The consequence of the growing deluge of data from the Internet of Things, mostly unstructured data, is a challenge for the worldwide networks. In an world where bandwidth demands continue to expand for a number of purposes, ranging from mobile data and video to 'bandwidth-intensive' applications such as video conferencing and ever larger files, the battle for the internet is on.

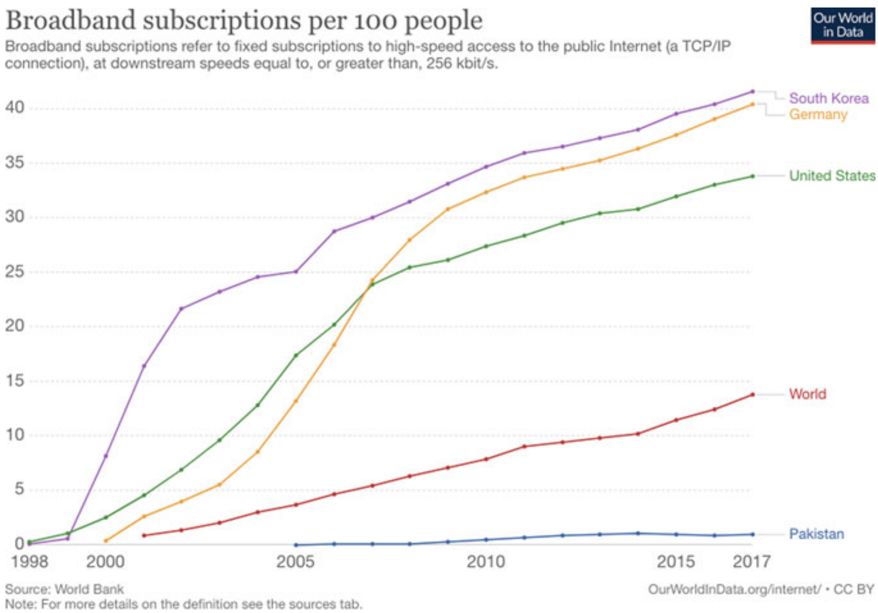


Fig. 5 Broadband subscription worldwide. *Source* World Bank

This is one of the reasons why, in the case of Wide Area Networks (WAN), you see a progressive shift towards hybrid networking with added sophistication to allow optimal use and scale of the available resources and a step towards Network Function Virtualization (NFV) and Software Defined Networking (SDN) in order to make networks more accessible, quicker and smarter (read: even more software based). Bringing millions and beyond (Fig. 5).

INetworks may also be broken down into classes depending on the amount of distances they offer.

PAN, LAN and WAN are mainly used by IoT devices. Once an IoT device is able to connect to WAN you are essentially home-free, the sky is the limit. But you need to get to the WAN and with a number of IoT devices this is always the challenge. All IoT devices that combine PAN with a smartphone will use your smartphone as their internet gateway (WAN).

The most popular networks in IOT are as follows:

(i) MQTT (Message Queue Telemetry Transport) is a lightweight protocol for transmitting simple data flows to middleware and applications from sensors.

The protocol operates on top of TCP/IP and consists of three components: client, author, and broker. The publisher compiles data and sends the data to subscribers. The broker tracks publishers and clients, checks their permission and guarantees protection.

It suits small, less expensive, low-power and low memory devices.

DDS (Data Delivery Service) is an IoT standard for machine-to-machine communication in real time, scalable and high performance. The Object Management Group (OMG) created this.

DDS can be deployed in low footprint devices as well as in the cloud.

The regular DDS has two major strata:

Data-Centric Publish-Subscribe (DCPS) which provides subscribers with the information.

(ii) AMQP (Advanced Message Queuing Protocol) is an application-layered middleware protocol for message oriented environments. That's an agreed international standard for this.

The processing chain for the protocol is composed of three components that obey certain rules.

Exchange—gets messages and puts them in queues
Message queue—store messages before customer app can safely process them
Information-Bound
Binding—notes the relationship between the components first and second.

(iii) Bluetooth is a short-range communication technology integrated into most smartphones and mobile devices, which for personal products, particularly wearables, represents a major advantage.

Bluetooth is well known for Smartphone devices. Yet not so long ago, the new necessary protocol for IoT devices—Bluetooth Low-Energy (BLE), or Bluetooth Smart, came up.

(iv) ZigBee 3.0 A low-power, low-data wireless network that is commonly used in industrial environments.

The Zigbee Alliance has also developed the common language for the Internet of Things—Dotdot—that allows smart objects to function securely on any network and to seamlessly understand each other.

- Standard: IEEE802.15.4 based ZigBee 3.0
- Frequency: 2.4 GHz
- Range: 10–100 m
- Data rate: 250 kbps.

(v) Wi-Fi is technology for device-wireless radio networking. It provides fast data transfer and is able to process large amounts of data. Inside LAN environments this is the most common form of connectivity.

1. Standard: IEEE 802.11 Basis
2. Frequencies: frequencies 2.4 and 5 GHz
3. Range: About 50 m
4. Service rates: 150–200 Mbps, max. 600 Mbps.

(vi) Cellular technology It the cornerstone of mobile networks. But it is also suitable for IoT apps which need to operate longer distances. Cellular networking systems such as GSM, 3G, 4G (and possibly 5G) can be exploited.

Huge volumes of data can be transmitted by the technology, but the power usage and expenses are also huge. So for projects that submit small amounts of information, it can be a great solution.

The Internet of Things has become the basis for digital transformation and automation, creating new business offerings and improving our way of living, working and entertainment.

Selecting the right type of connectivity is an unavoidable part of any IoT project.

4 Augmented Intelligence

Effective human-computer symbiosis, often called Augmented Intelligence, has the potential to successfully address some of these emerging challenges, perhaps more so than pure AI, at least in the foreseeable future. Nevertheless, without the ability to capture, compile and analyze the huge amount of data generated by IoT devices, either by AI alone or assisted by human intelligence, the promise of IoT does not materialise.

With all the data that is gathered, information or insights can be derived only fro data analysis. There are certain technologies which are cognitive which enable this. These cognitive technologies and its accompanying models are collectively known as augmented intelligence. This in a way automated intelligence and eliminates human intervention. But with human intervention it can be made better. The amount of data generated is like a heap of mountain. The role of analytics can be said like searching a needle in a haystack. The relevant data may be miniscule to help in decision-making.

Descriptive, predictive and prescriptive analytics are ways of coming out with better decision guidelines. In descriptive analytics various data visualization tools are used like Tableau etc. These tools make it easy to understand large data sets which otherwise seem incomprehensible. Predictive analytics, as the name suggests, predicts what can happen in future based on the historical or past data. Predictive analytics usually use machine learning. This however, totally depends on human intervention for taking decision on future action. Prescriptive analytics develops models for best course of action to be taken depending on the analysis done using existing data.

Most of the image-processing activities are done using computer vision technology. This technology is mostly used in medical imaging and prediction of certain diseases and thereby, treatment of the same.

Natural Language Processing or NLP is used for working with text the way humans do. It has various techniques and it results in a string of characters. It identifies data based on text. For example NLP is used to distinguish between spam mails from real ones.

Machine learning, natural language processing, data analytics and other AI applications, with large-scale data and without human judgment, can work extremely well. In contrast, areas such as creative marketing advertising design, personal sales, hiring and mentoring of employees, strategic decision-making and disease treatment are prime examples of human judgment, managerial intuition and human-computer symbiosis can still enhance work performance. Human intelligence is still needed to make decisions about human-oriented data related to employees, customers and partners that include their preferences, past behaviors, habits, emotions and personalized information. In general, human beings outperform computers when dealing with uncertainty, vagueness and incomplete knowledge and when requiring emotional intelligence and judgement, elements that are still considered to be the most important weakness of AI. Augmented Intelligence incorporates human special abilities that can not be reproduced by AI This has been confirmed by many visionaries like Bill Gates. Elon Musk etc. Large-scale IoT problems can always not be solved exclusively by machines or humans. In IoT implementations, therefore, there are major possibilities, combined with the notion of Augmented Intelligence. Machine learning methods and human intelligence should be combined to retain absolute influence of mankind.

Speech recognition is yet another augmented intelligence technique. It works on transcribing human speech. There are many challenges in this as accent differences, background noise may be there, the speed in which a person speaks differs from another. It is used now for placing orders by customers in online business.

All the above techniques are adopted by IOT is because there is a huge reservoir of data and the urgency of information is driving organizations to adopt real time data analytical tools. Large data sets help in improving artificial intelligence models.

There are also various challenges in augmented intelligence. Flaws in data or lack of data leads to inaccurate analysis. Since IOT and businesses generate unstructured data it is difficult for legacy systems to analyze data as they are suited for structured data. All the IOT-driven systems need applications that can handle unstructured data. Also real-time processing cannot be done on legacy systems (Fig. 6).

Symbiosis between human and machine signifies the mutual relationship between human beings and computers. A symbiotic human-computer relationship will emerge in the context of IoT when IoT collects the data and AI tools perform the routine data calculations based on human-determined criteria and prepare the insights for evaluation and decision making. The fundamental concept of human-computer symbiosis is that computers and humans have complementary abilities and the capacity to solve problems. Intelligence is increased by combining computer and IoT computing capacity with human being's intellect, intuition, and "common sense".

Augmented Intelligence is not a new term, so it is related to Human Computer Interaction Research (HCI), but IoT is a new area for HCI research. Interestingly, researchers at HCI have always warned of the simplistic nature of AI, which views humans as perfectly rational machines. Instead, HCI aims to strengthen the symbiosis between the interpretive and the emotional. Most approaches to Increased Intelligence rely on crowdsourced methods and gamification in the context of conventional HCI. They are able to provide Augmented Intelligence interface solutions which can be applied to IoT devices to improve human-IoT interaction design (Fig. 7).

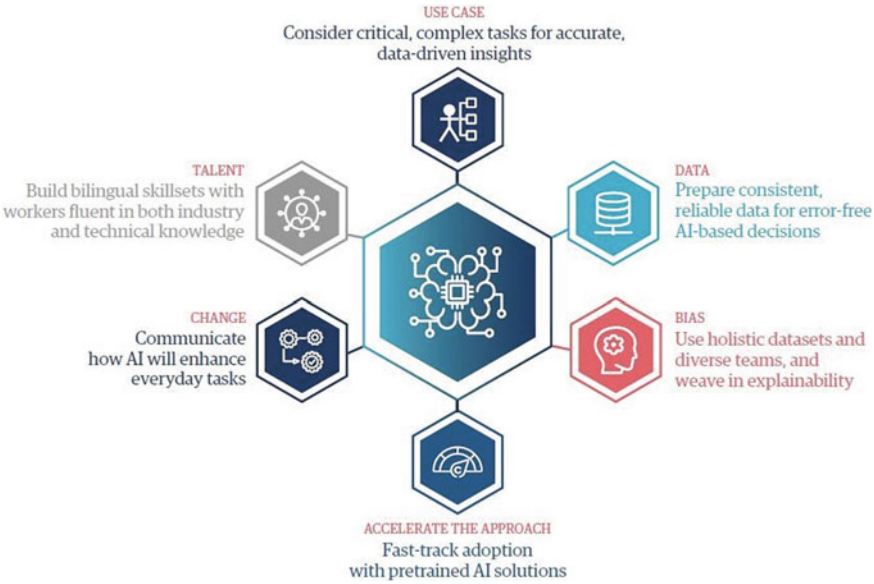


Fig. 6 How augmented intelligence is changing finance and accounting. *Source* GenPact

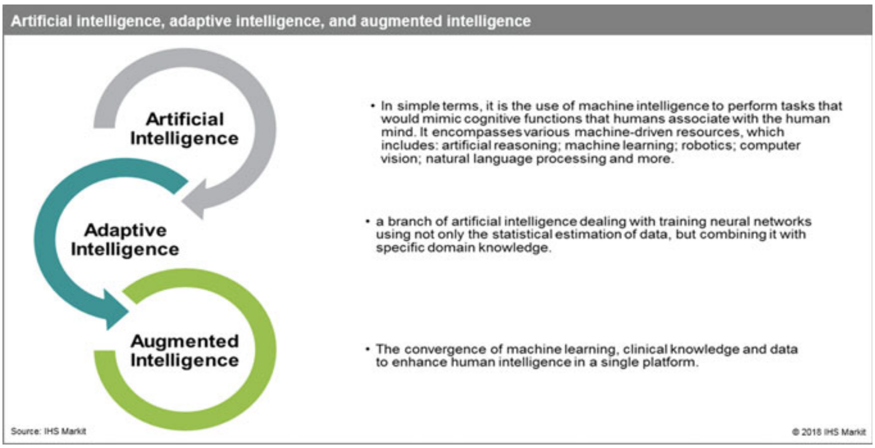


Fig. 7 Artificial intelligence, adaptive intelligence and augmented intelligence. *Source* IHS

5 Augmented Behavior

After sensor does its work and analysis of data is done, it leads to some action. This action is termed as augmented behavior. This action further leads more data and the cycle starts once again. It is very difficult to separate augmented intelligence from augmented behavior. One can say that augmented behavior is action which can be observed.

Machine-to-Machine (M2M) is a concept by which technologies allow machines to communicate wit each other. This communication leads to some kind of action. In today's time it is used interchangeably with IOT.

Machine-to-Human interfaces can be termed as actions suggested by machine to humans based on certain data. However, it is the discretion of the individual to take an action.

What organisations do is to include M2M as well as M2H as organisations need both. For them information creates values in different ways. However, the entire process requires improvements in the machine's cognitive abilities and execution of action based on the data. This action has to be very accurate. If any flaw is there, it will lead to the downfall of the entire process. This is especially true when organisations start using robots. These robots when malfunction can lead to disaster.

Why organisations have started using robots is that over the years the functionality of the robots has become so much better that errors are very few. Moreover, their cost is constantly decreasing. The entire technology used is becoming cheaper (Fig. 8).

In the above graph it can be seen that supply of robots in the year 2021 is supposed to increase by 14% annually.

The progress in algorithms has been a motivator for shifting routine works done by robots to more sophisticated tasks. In most advanced robotics demands more data than may be available for increasing functionality of robots.

Controlling human behavior through robots is a very difficult task. Moreover, it is considered that machines are more reliable than human. But in terms of data we can say that any flaw can lead to malfunction and major disaster. Privacy is a also a major concern amongst such machine. Alexa is a part of IOT. But if we see its functioning,

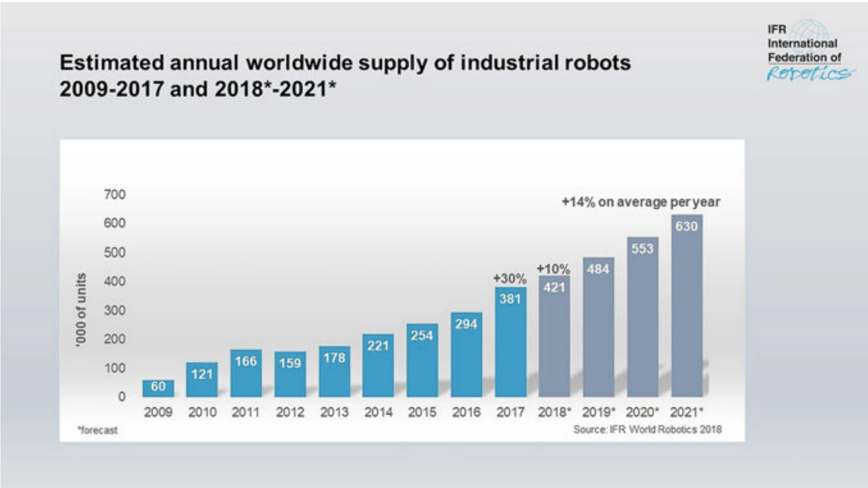


Fig. 8 Estimated supply of robots. *Source* International Federation of Robotics

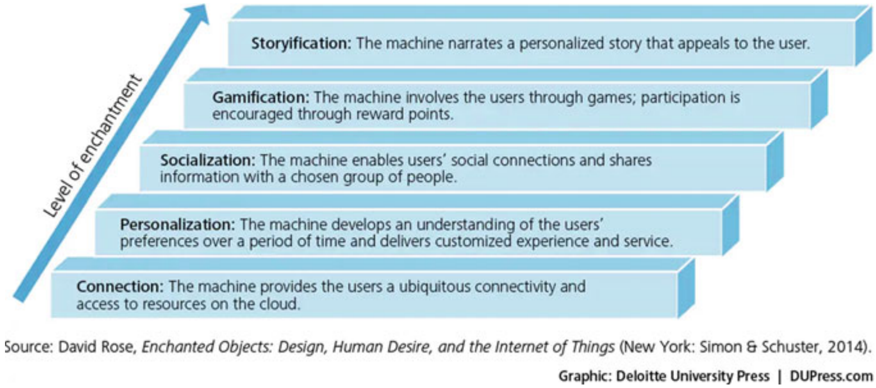


Fig. 9 Machine enchantment hierarchy. *Source* Deloitte University Press

its microphone is always on because if it is not on how will it know when somebody calls out for her. This means in a way Alexa 'is listening to everything we are talking' Can you imagine the risk? This can lead to breach of privacy and can lead to major security concerns.

Figure 9 shows the level of enchantment in IOT. The hierarchy can be clearly explained with the complexity in technology.

6 Standards

The first Global Net-connected computers appeared in the 1980s. Coca-Cola developed its vending machine, which monitored the temperature of a storage place and kept track of the amount of bottles therein. Kevin Ashton, an RFID-technology academic, is considered to formulate the word "Internet of Things" in 1999.

In the first decade of 2000 a rapid growth in the technologies of IOT was seen. At this time many applications became popular. Using these applications many organisations started divulging in small as well as large projects like automatic cars, trackers for health and fitness etc.

The IoT has developed into an entirely "new Internet," meaning that not all standard networks have been able to satisfy their needs and provide seamless communication. That's why developing professional IoT communication protocols and standards became a critical necessity. However, the Internet of Things also uses some of the existing technologies (e.g. HTTP).

An IoT framework has three-tiered architecture: computers, gateways, and data systems. The data travel between those rates across four different types of transmission channels.

1. Device to Device (D2D)—Direct contact between two smart devices, if they share information instantly without intermediaries. For example, industrial

robots and sensors are linked directly to each other in order to coordinate their activities and perform part assembly more effectively. This form of communication is not yet very widespread as most devices are unable to handle such processes.

2. Gateway system-telecommunication between sensors and gateway nodes. Gateways are the more powerful electronic devices than sensors. They have two key functions: consolidating sensor data and sending it to the appropriate data system; analyzing data and returning it to the machine if there are any issues. Depending on the gateway computing capacity, there are various IoT gateway protocols that might be better suited to this or that solution.
3. Portal to data systems—Shift of data from server to the respective data network. To decide which protocol to use (frequency of burst and congestion, safety criteria and how many parallel connections are required) you will analyze data traffic.
4. Between data-systems—Data center or Knowledge Sharing Cloud. Deploying and integrating protocols for this form of communication with existing apps should be simple, have high availability, reliability and fast recuperation of disasters.

A variety of activities takes place in IOT like accumulation of data, handling and processing of it and also to some extent storage of the data. This entire data is then collected and made comprehensible. One can say this data is aggregated and is then available for analysis. But it must be kept in mind that the aggregation of data must be according to certain guidelines. These guidelines are termed as standards.

Aggregation of data has to maintain two important standards, that of network protocol and that of communication protocol. Certain regulatory standards are also to be maintained. These standards are basically for security and privacy. IOT cannot function without these standards.

Some of the regulatory standards are use limitation and purpose. This standard states that whichever entity collects data, has to state in full the purpose of collection of data and the scope of usage of the data. This standard puts a limit of usage of data by IOT. The next regulatory is that of choice and notice. The entities collecting data have to inform the user before collecting such data and bring to their notice if any personal information is being shared. The principle of data minimization says that an organization should collect data only for the specified purpose and delete it after its use. Another regulatory states that the entities that collect data are accountable and have to ensure that no unauthorized person accesses the data or modifies it.

7 Conclusion

IOT is still in a nascent state. It still streamlining and organizing itself. IOT has already become a part and parcel of every organization and every individual. Knowing that we are really only at the beginning of the evolution of the Internet of Things, with the possible exception of certain industries, this in turn is good news for many ICT companies and infrastructure providers. In the area of consumer applications, the IoT is still pretty similar to everywhere today.

Recent advances in sensors and networking technologies have led to substantial growth in devices such as wearables on the Internet of Things (IoT). They are focused on technologies of Artificial Intelligence (AI), such as machine learning, deep learning, computer vision, processing of natural language and analytics of big data. These advances led to the development of autonomous systems, for example smart home, smart cities etc.

Could these intelligent IoT devices make redundant all human workforce? In the future of the job, would humans become meaningless? Will AI systems completely replace human beings, or will the IoT rather enhance human intelligence and empower us to solve the most pressing issues of our time in a more satisfying way than we'd ever dreamed about?

Effective human-computer symbiosis, often called Augmented Intelligence, has the potential to successfully address some of these emerging challenges, perhaps more so than pure AI, at least in the foreseeable future.

Machines alone could dominate decision taking in most applications in some distant future. Nevertheless, there will be a relatively long interval, during which human beings and computers working together in intimate partnership will make the big intellectual advances. IoT is such an area that the human and computer intelligence can be integrated to solve emerging problems. In addition, we need to adapt our culture to the wider social, cultural, behavioral and ethical consequences of AI and Augmented Intelligence. Both have an impact on the future of employment, corporate competitiveness, blurring market borders and other activities related to legal, policy and governance. For the time being, suitable IoT designs would retain a fair degree of human control and supervision and give the human race a chance to get to know how to assign power to machines.

References

1. Baker, K.: The Digital Revolution Chairman of the Edge Foundation (2018)
2. Bojanova, I.: The digital revolution: what's on the horizon. IT Prof. **16**(1), 8–12 (2014). https://doi.org/10.1109/MITP.2014.11
3. French, A.M., Shim, J.P.: The digital revolution: internet of things, 5G, and beyond. Commun. Assoc. Inf. Syst. **38**(1), 840–850 (2016). https://doi.org/10.17705/1CAIS.03840
4. Mühleisen, M.: The long and short of the digital revolution. Finance Dev. **55**, 4–8 (2018). https://www.researchgate.net/publication/329515181_The_Long_and_Short_of_The_Digital_Revolution
5. Wyman, O.: Digital Revolution (2014)
6. Zysman, J., Kenney, M.: The next phase in the digital revolution: intelligent tools, platforms, growth, employment. Commun. ACM **61**(2), 54–63 (2018). https://doi.org/10.1145/3173550

Dr. Seema Sahai is Associate Professor in IT & Operations at Amity International Business School, Amity University Noida. She is a Ph.D. in Management and has a journey of 23 years in academic and consistently striving to create a challenging and engaging learning environment She has to her credit a number of publications in reputed national and international journals accompanied with participation in conferences. She is currently the HOD of IT & Operations department at Amity International Business School. She has a corporate experience of 2 years and has many projects to her credit.

Dr. Richa Goel is Assistant Professor-Economics and International Business at Amity International Business School, Amity University Noida. She is a Ph.D. in Management and has a journey of almost 18 years in academic and consistently striving to create a challenging and engaging learning environment where students become life-long scholars and learners. She has to her credit a number of publications in reputed national and international journals accompanied with participation in conferences. She is serving as a member of review committee for conferences journals and acting as Lead Editor of Annual International Referred Journal and Research Coordinator with Amity International Business School. Her area of interest includes Economics, Business Law, Human Resource Management and Diversity Management.

Prof. (Dr.) Gurinder Singh Group Vice Chancellor, Amity Universities, Director General, Amity Group of Institutions and Vice Chairman, Global Foundation for Learning Excellence, has an extensive experience of more than 21 years in Institutional Building, Teaching, Consultancy, Research & Industry. He holds a prestigious Doctorate in the area along with a Post Graduate degree from Indian Institute of Foreign Trade where he illustriously topped with 7 merits. He holds the distinction of being the youngest Founder Pro Vice Chancellor of Amity University for two terms, the Founder Director General of Amity International Business School and the Founder CEO of Association of International Business School, London.

Applicability of Machine Learning Algorithms for Intelligent Farming

Bharti Verma, Nikhil Sharma, Ila Kaushik, and Bharat Bhushan

Abstract Agriculture contributes enormously to the growth and economy of a country due to which it becomes important to upgrade the agricultural facilities for farmers that simulate them for cultivating good quality crops with high production rates. This paper sets sights on classifying different types of crops grown, and predicts which crop is best suited for particular location for boosting the production factor. Further, this ML model will be integrated with Internet of Things (IoT) to build an intelligent irrigation system that itself decides whether the crop-land needs to be irrigated or not. This system uses decision tree algorithm, Arduino, sensors, and bolt IoT kit. By means of feature extraction and data analysis techniques, we were able to select highly meaningful and best contributing variables from gathered data that were affecting the prediction values. Also, we discovered and unleashed the working statistics behind certain powerful ML algorithms. Strong statistics like hypothesis testing, chi-square testing and Euclidean distance are thoroughly discussed. Different classification models like K-NN, decision tree, SVM (Support Vector Machine) and logistic regression were implemented and compared in order to reach the best suited model for forecasting the crop class label.

Keywords Chi-square test · Sigmoid function · Entropy · K-NN · Decision trees · Data analysis · IoT

B. Verma
HMR Institute of Technology & Management, Delhi, India

N. Sharma (✉)
Department of Computer Science & Engineering, Delhi Technological University, Delhi, India

I. Kaushik
Krishna Institute of Engineering & Technology, Ghaziabad, India

B. Bhushan
School of Engineering and Technology, Sharda University, Greater Noida, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_6

1 Introduction

Machine Learning (ML) is an incredibly exciting technology that provides computers the power to learn and react. From this methodology, machine gains ability to learn thus acquire knowledge to solve problems just like humans do [1]. It relies on mathematical calculations and statistics to formulate its algorithms and models. These models are used in wide range of applications like computer vision, email filtering and many more [2, 3]. ML is considered highly essential where it is inconvenient to produce conventional algorithm for effectuating provided task. With advancement in technology use of artificial intelligence in agriculture is growing rapidly with time. As agriculture enacts largely in a country's economy and due to increasing population there's a constant pressure on farming sector to grow crops with high productivity [4]. Thus, ML is used widely in this sector to boost crop yields by predicting what conditions and which crop in a particular land can produce effective results [5]. Also, machine learning is applied for identifying bug hunters using computer vision and further providing preventive steps a farmer could take using a mobile application [6]. This chapter comprehends data analysis techniques and distinct classification models from the field of machine learning and by making use of both the above mentioned approaches a ML model is formulated that successfully was able to classify the crop type that should be grown in a particular area in order to improve the production factor [7], this was further integrated with IoT to construct an intelligent irrigation system for farmers [8], this system does not depend on the farmers knowledge or experience to irrigate the field [9], it itself is capable to comprehend if the crops need more water or not by analyzing the condition of the soil. Subsequently, valid course of action is taken by the system [10]. In today's industry, where large amount of data is being produced on regular basis there is a need for properly collecting, cleaning and administering that data, consequently, that is something what we do in data analysis techniques [11]. It's essential to analyses raw data before feeding it to the any machine learning or neural network model as only clean and meaningful data will dominate in producing successful requisite predictions [12]. By analyzing and examining raw data we ought to extract useful information. Data has become utmost important entity that has evolved the IT industry trends and approaches for handling different problems [13]. There are diverse technologies available for interpreting data [14]. Majorly, various techniques available are diagnostic analysis, statistical analysis, predictive analysis, text analysis and prescriptive analysis [15]. The data collected must be organized into readable context for example, it may be in structured format that is in form of table, possessing table rows and different features as columns, that is the most commonly used format all over the world [16]. In practical, the data obtained from different sources is not error free it may contain missing values and duplicate data within cells, for data to be error free it must be cleaned and null values should be handled beforehand [17]. There are various methods present we can use for cleaning of data. Although, selection of cleaning methods solely depends upon the type of data in use for example, incorrectly entered values in data could be handled by quantitative analysis practices [18]. Analyzing data could be briefed as filtering of raw values and holding data that is contributing meaningful insights. Also, Robust statistics like hypothesis testing, chi-square testing and Euclidean distance are thoroughly discussed. These are some of the underlining techniques working under the hood of ML algorithms [6]. This paper also covers the thorough explanation about working of some classification algorithms like SVM [19], K-NN [20], decision trees and logistic regression.

The rest of the chapter is organized as follows. Section 2 comprises of literature review. In Sect. 3, we will discuss about data analysis methods and feature engineering. Section 4 describes the mathematics used for classification algorithms like information gain, chi-square testing, Euclidian distance and more. Section 5 illustrated about the working statistics of classification algorithms. Section 6 covers implementation of the processes discussed above as well as the integration of IoT with ML models followed by the conclusion and future scopes of the chapter in Sect. 7.

2 Literature Review

Ravishankar et al. [21] stated the fact that many farmers involved in agricultural field does not have much educational background and are illiterates and that is a stumbling block for agriculture due to the unawareness of new technology. The ongoing agricultural exercises are not prudentially successful also in India production of major agricultural products are precarious. Botsiou et al. [22] have suggested the appropriateness of using Information and Communication Technology in agriculture. Preliminary small-scale study was presented in their research. Song et al. [23] explored uneven two-class unsupervised classification problem of satellite images. The study focused on mapping of CRP, USDA programs in order to conserve and enhance water, soil and resources belonging to wildlife. In 2016, Mohanty et al. [24] stated that although, security of food is mostly threatened by plant diseases even then these diseases remain undetected in most fraction of the world that is mainly a consequence of unavailability of requisite infrastructure though, due to increase in advancement in day-to-day machinery and computer technologies has led a way for recognition of these diseases. The model predicts with 99.35% accuracy. Singh et al. [10] stated that high clarity images and sensor data of crops is an outcome of advancement in high throughput and automated images, this set of information can be used for extracting features and patterns that further need to avail machine learning techniques for fully understanding those information and patterns. Those machine learning techniques are availed in different steps like classification, identification, quantification and prediction. Kumar et al. [25] states that how big data analysis is leading a way for intensifying distribution channel. In their research have scrutinize the discussion about major difficulties in cultivation decision making also presented detailed analyzer report regarding the present status of market, weather, storage capacity, soil and instantaneous request of the market. Huang et al. [26] discusses data processing platforms for intelligent agriculture also had a detailed study about distinct processing classes like Hadoop, Cassandra and storm stream they tried to make stand out the idea of how the phase of production can benefit the agricultural stakes. Varman et al. [27] projected a smart irrigation system based on hybrid approach consisting of different sensors, that were used to gather remote data. The system incorporates developing WSN that integrates different methods of IoT, deep learning and remote sensing together. Payero et al. [28] established a model incorporating mobile computing, IoT (Internet of Things) and big data, in which the farmer retrieved the information regarding fertilizer requirements of the crop-land by its soil samples. The farmer in turn is notified with needed extent of fertilizers in field via mobile application. Byakatonda et al. [29] proposed deep learning ANN model that predicts yields of maize crop based on various factors such as precipitation length and climatic indices. In 2019, Nevavuori et al. [30] proposed that Artificial intelligence has proffered noticeable remunerative results for agricultural practices and also for problems like pattern recognition which is similar to crop yield prediction.

Yang et al. [31] suggested Convolution neural network model that forecast rice crop production. Cynthia et al. [32] discussed that another area of precision in agriculture is to determine the area under influence of diseases and pests, regarding this issue a R-CNN deep learning model was proposed and area under disease influence was localized. In their work, Samsonovich et al. [33] discussed the Colossal improvement in domain of machine learning and artificial intelligence have boundless capable results. The spatial factors regarding crop yield are learned by the neural network from RGB image of high resolution Long short-term memory-based model was proposed by Hang et al. [34] which was used to detect disease in potato plants. Shafi et al. [35] Projected integrated model employing machine learning and IoT that produces heterogenous data using drone and possesses varying temporal fidelity. Elavarasan et al. [36] Proposed a model drawn upon a deep learning Recurrent Neural Network over a reinforcement learning algorithm in order to predict yield of crops in a farm.

3 Data Analysis and Its Techniques Used in Machine Learning

In modern world data have plentiful information in form of high dimensional datasets. Though, all that information may or may not be useful for us depending upon different situations. In data science it's very important to regularly check and update the data. Also, this data should be filtered and serialized for further use in machine learning models to pull off lucrative results. For this we frequently use feature selection and feature extraction techniques also known as feature engineering in machine learning, some of these techniques are explained as follows.

3.1 Feature Selection

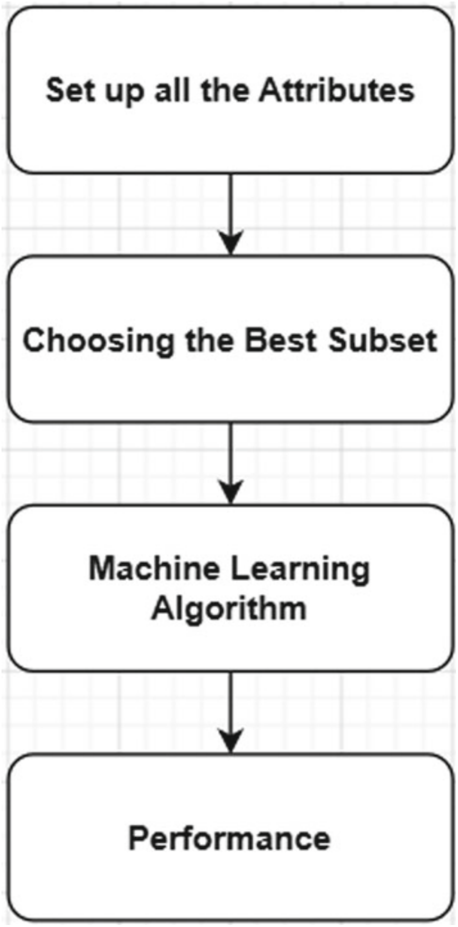
Feature selection, also addressed as variable selection or subset selection. Where we intend to pick out such features of data which highly influence our target feature. This is considered as a dominant technique in Data mining community. It is mostly used when there is high dimensional data, also feature extraction and selection are very frequently applied in Machine Learning where large group of features are analyzed and a subset of certain features are nominated which favors the application of selected algorithm [37]. The best subset is the one which contains least number of dimensions that must pitch in maximum accuracy. Feature-based feature selection methods involves evaluating the associativity between each target- feature value and the input-feature value using statistics. Electing those provided input variables that has strongest relationship with the target feature. These methods are effective and fast although, the choice of statistical measures is based on the data type of both the input and output variables. Feature selection has been active research region in statistics, data mining and pattern recognition communities. The main concern sector of this technology is to select a subset of input values by banishing features with low or no predictive information [38]. These feature selection methods simplify the machine learning models by making them easier to interpret and implement by the researchers. Besides, this methodology enhances the generalization within the model that narrows the problem of overfitting. Hence it improves performance for forecasting, helps in better understanding of data, also contracts the computational space as well as time which is required to run the algorithm. These methods of feature selection are broadly classified into three classes namely Filter methods, Wrapper methods and embedded methods.

3.1.1 Filter Method

This approach avails the attribute ranking technique so as to nominate the features for arranging them. Here, the nomination of variables does not depend upon the type of classifiers used. Whereas, the ranking expresses how important and meaningful each and every feature is and which of them more likely accounts for categorization it generally picks the subset of features as a step before performing main processing that is individualistic of the selected predictor [39]. In this methodology for eliminating the irrelevant features the ranking method are applied before the classification. The feature selection exercise is carried out as a step of pre-processing that accommodates no induction algorithm below are the few example of filter methodology.

- *Chi-Square Test*: much normally, this course of action is availed to validate the dependency levels of two events. If a data is provided for only two events, we could obtain the expected count and observed count and in turn this will measure to what extent both calculated counts are inferential with one another.

Fig. 1 Filter-based method



- *Variance Threshold*: This method selection vanishes each and every attribute those variances do not meet a threshold. In general, it abolishes all zero-variance attributes that means every attribute that gives the similar value in every sample.
- *Information Gain*: Information gain technique assesses what amount of information a variable provides regarding the class. Thus, one could elect which variable among the proffered training set of variables is utmost useful for recognizing a distinction between the categories to be studied. Figure 1 shows the Filter-Based Method in Machine Learning.

3.1.2 Wrapper Method

The Wrapper technology was made popular by researchers George H. John and Ron Kohavi in 1997 [40]. Here, subsets of features are made based upon their predictive

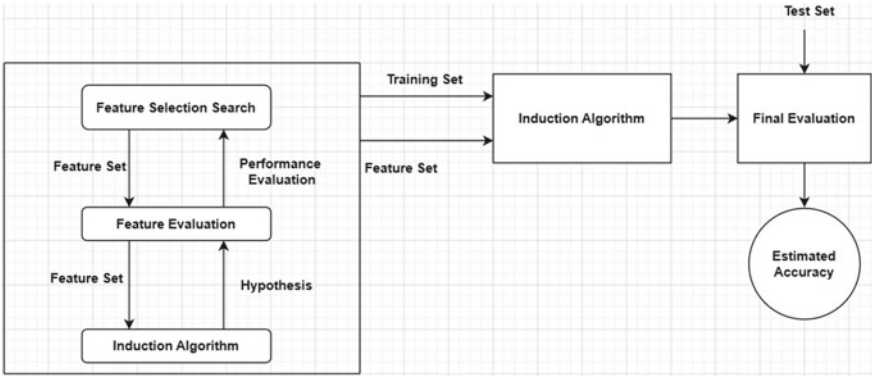


Fig. 2 Wrapper feature method

powers, by considering the learning machine of concern as a black box. In Fig. 2, a set of training instances are used to depict the induction algorithm in which every instance is specified by class labels and vector features values. The induction algorithm also referred as the black box used to forge the classifier which in turn is used for classification. in this technique, induction algorithm contains subset selection algorithm as a wrapper around it [41]. Drawbacks of this methodology includes the mass computation that is required to develop the feature subset. Few examples of wrapper approach are broached bellow:

- *Genetic Algorithms*: the aim of this algorithm is to perceive a small set of variables. CHCGA is another similar improved design of the algorithm that merges faster and proffers a more effectual search by preserving the distinctiveness and eluding the doldrum of the population.
- *Recursive Feature Elimination*: This is the type of process that fits a model and eliminates the fragile variable till the requisite number of variables are satisfied. The model's coefficient or feature important attribute is taken in consideration while ranking the features.
- *Sequential Feature Selection*: This is a naive model that begins with an empty set. A single feature is added to the initial step that represents the largest gain for the objective function, from the following step on the left-over variables are appended separately to the present subset and as a consequence the new subset is assessed till the requisite number of variables are appended, this process is repeated.

3.1.3 Embedded Features

Embedded method seeks to unite the efficiency of both the earlier methods. Here, the procedure of training takes place by nomination of variables that are mainly specific to provided learning machine. Basically, the method learns which attribute will contribute greater to the accuracy of learning model [42]. Below are some of the Embedded based feature methods.

- L1 Regularization methods like LASSO affirmation for Least Absolute Shrinkage and Selection Operator which is a linear model that gets the measure of sparse coefficients it is also considered useful in some context because of its ability to proffer solution with lesser parameter values.
- Ridge Regression (L2 Regularization): this algorithm is also referred as Ridge Regression also sometimes known as Tikhonov Regularization in this the loss function is the regularization and the least square function which is used to solve a regression model.
- Elastic Net: L1 and L2 as regularizes are used to train this linear regression model which simulates for learning a sparse model where few of the weights are non-zero such as Lasso and on contrary maintaining the regularization properties of Ridge.

4 Background Statistics of ML Algorithms

Here we will discuss some of utmost powerful statistical methods that are being used behind the scenes of these feature-based methods. These are core probability and statistical approaches that helps us in evaluating the relationships amongst the provided data. Some important techniques of all times are mentioned below.

4.1 Chi-Square Statistics

The χ^2 statistics is the customary test that evaluates the divergence from the distribution expected a condition that one presumes, the feature occurrence does not depend upon the class value. This method assesses the non-dependency between the category and a term [43]. It becomes irregular and unpredictable when comes to small expected count. Which is something general in test classification because of two main reasons firstly, having rarely arising word features and second, because possession of only small amount of positive training samples for a method. The occurrence of the term and occurrence of class are considered as two events in feature selection [44]. The χ^2 statistics is comparable to using of hypothesis testing on a given distribution because it associates with the values of feature in question. There are some variations in chi square statistic usage of which depends upon how one collects the informational data and what type of hypothesis is being examined however, the variations follow the same idea that is to collate the expected values with the values one actually gathers [45]. The most common form used is contingency tables. The formula used in the χ^2 statistics is

$$c_c^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i \tag{1}$$

Here, the subscript ‘c’ depicts the degrees of freedom, ‘E’ represents the expected value and the observed value is ‘O’. Whereas, ‘i’ is the ith value of the contingency table. Higher the value of chi squares the higher will be the correlations between the data that means the count of categorical variable are closer to what one would expect [46].

4.2 Euclidean Distance

It is the most general form of calculating distance. Basically, it is the distance between two different points in Euclidean space. Euclidean geometry is precisely used for the spaces of two and three dimensions. However, it can smoothly be generalized for higher dimension spaces. This method scrutinizes the distance as a function of “root of square differences” between set of coordinates of the data [47]. Euclidean distance is a fancy name for distance formula we learned in high schools [48]. However, in terms of machine learning it is used for measuring the ‘similarity’ or we can say the correlations between two vectors. Formula for calculating Euclidean distance is given below, if p, q ∈ R: where p and q are two arbitrary points in space belonging to set of real numbers.

$$Dis(p, q) = \sum_{i=1}^{n} \left((p_i)^2 - (q_i)^2\right)^{1/2} \tag{2}$$

These Euclidean distances are generally evaluated from raw data and not from standardized one. One of the advantages of this process is addition of new vectors does not affect the distance between any two coordinates [49].

However, the distances may get highly affected by distinctness in scales admits the dimensions from where they are being evaluated [50, 51]. For example, if there is a dimension representing measured capacity of water and that is being converted into liter from kilo-liters by you, then this is supposed to produce out of order results of our cluster analysis.

5 Classification Algorithms Used

In machine learning, there are certain specific algorithms used for solving different class of problems such as regression, classification and clustering algorithms. Classification algorithms are the supervised learning algorithms that puts an overview of different classification methods. The classification algorithms used in this study

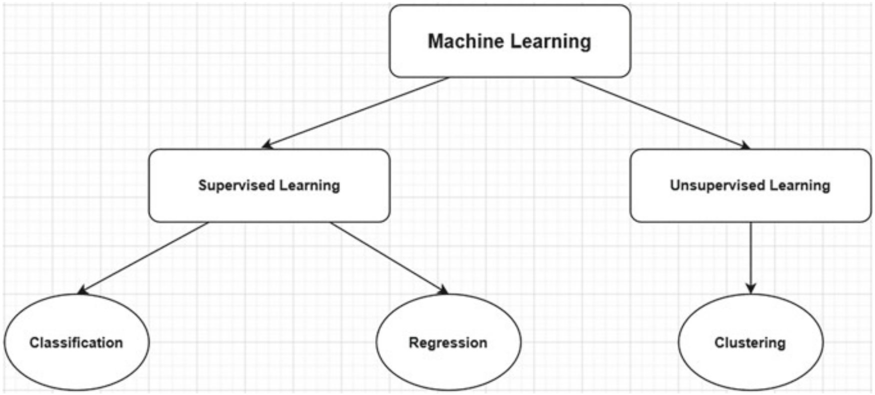


Fig. 3 Types of algorithms

are explained below with their working principals. Figure 3 shows different types of algorithms in Machine Learning.

Specially Those methods that are commonly used in data mining techniques with different principles. Classification is a technique that categorizes data in different number of unique classes and in turn labels are assigned to each class [52]. The main objective of classification is to identify the class to launch new vector by analysis of training set by visualizing proper parameters. In more general sense predicting the target class using above process is referred to as classification.

5.1 K-Nearest Neighbor's

K-NN is one of the most commonly used supervised learning algo of machine learning. In area of data science where we continuously build classification and regression model k-NN is considered as a powerful algorithm as it could be used both in regression and classification problems however, it is mostly used in classification problems. There are three main parameters that comes into consideration while assessing any of the machine learning algorithms, these are first, ease to interpret output second, calculation time and last, predictive power and k-NN fairs across all the three parameters of aspect, mostly this algorithm is used for its ease of interpreting output and low calculation time. k-nearest neighbors use 'feature similarity' for predicting the values of new datapoints that is, the new data coordinate in space will be assigned a weight based upon how contagiously it matches the points in the training set, it uses Euclidean distance statistics in background to calculate for the nearest neighbors of the provided data point in space. In k-NN 'k' is represents the value of nearest neighbors that is the maximum number of nearest data points taken into consideration while plotting the new data. However, it is not an easy task to find an optimal value for variable 'k'. If we choose a small value of k then the noise would have a greater influence on the result and if we choose a larger value it may become computationally expensive as a consequence [53]. Although data scientists usually choose its value for an odd number if the number of classes is two. Another simple approach for selecting the value of k is using the formula, k = sqrt(n), where 'n' is number of datapoints. k-NN can be simply be understood by an example, suppose the aim of the problem is to distinguish between two geometrical shapes namely rectangles and hexagons, the new data point placed in space will be distinguished as a hexagon or a rectangle on the basis of which shape's data points are the nearest neighbor of newly added object. As here a pentagon is the new appended data point. Here, the pentagon is more similar to the shape hexagon rather a rectangle because it possesses greater number of hexagonal data point nearest to it rather than rectangular points.

5.2 Support-Vector Machine (SVM)

SVM also known as support vector networks is a powerful yet flexible supervised learning algorithm, they have their unique way of execution in contrast to other ML algorithms, for a dataset consisting of a feature set and data set, an SVM classifier will generate a model for predicting categories for new data point. Lately, they have outgrown popularity because they can handle many categorical and continuous values. It basically depicts various classes in hyperplane in a multidimensional space. SVM uses iterative approach to generate a hyperplane in order to minimize the error. A maximum marginal hyperplane is detected after separating the classes of the proffered dataset. A margin is defined as the space between two lines on the nearest data points of two classes it can be assessed as perpendicular distance from the dividing hyperplane to the support vectors. Hyperplane is a space or plane that is dividing set of objects occupying distinct classes. Whereas, support vectors are datapoints that are close to the hyperplane. In real, implementation of this algorithm requires a kernel that converts input data plane into desired form here, a high dimensional space is being created by using the input that is a low dimensional space by applying the technique known as kernel trick in simpler form it converts non-separable problems into separable ones by adding new dimensions [54]. However, the mostly used type of kernels are liner and polynomial kernel. In linear kernel dot product between any two observations is used.

$$\text{For linear kernel:} \mathrm{K}(\mathrm{x}, x_i) = \mathrm{sum}(\mathrm{x} \times x_i) \tag{3}$$

This implies that product of any two vectors say x and x_i is the sum of multiplication of each set of variables. polynomial kernel is more general version of linear kernel. It is used for non-linear input space.

$$\text{For polynomial kernel:} \mathrm{K}(\mathrm{x}, x_i) = 1 + \mathrm{sum}(\mathrm{x} \times x_i)\hat{}\mathrm{d} \tag{4}$$

where d represents degree of polynomial in practice the degree is provided manually in learning model. SVM is suitable and goes well with data having high dimensional space.

5.3 Logistic Regression

Logistic regression is yet another technique from field of statistics applied in machine learning. By estimating probabilities using a logistic function logistic regression computes the correlations between the categorial dependent vector and one or more non-dependent vectors. Before diving into the parameters of logistic regression first understand what are the calculations that are taking place behind this algorithm. The linear part of the algorithm or in simple terms the weighted sum of inputs evaluates the log of odds or log-odds of a successful event, in practical, for all observed values or at each level log-odds for input variables is estimated [55]. Let's first understand what are odds, given a probability of success (p) that the regression model forecasts one can transfigure it into odds of success as simply dividing the probability of success by probability of no success.

$$\text{Odds of success} = \mathrm{p}/(1-\mathrm{p}) \tag{5}$$

We evaluate the logarithm of odds more particularly the natural logarithm or log base-e. This value is mentioned as the log-odds and can be introduced to as the logit (logistic unit), a unit of measure.

$$\log \text{ of odds} = \log(p/(1-p)) \tag{6}$$

This log of odds can be converted back to probability of success by assessing the exponential of log odds. We can simplify the so far evaluation as follows:

$$P = 1/(1+\exp(-\log \text{ of odds})) \tag{7}$$

Let's take, z = log of odds, then the equation will be simplified into:

$$P = 1/(1+\exp(-z)) \tag{8}$$

This equation above is well known as the sigmoid function or a logistic function thereby ensuring the value of probability function remains between 0 and 1. This sigmoid function contains values very close to 0 and 1 across its domain this fact makes it very much suitable for application in classification process.

5.4 Decision Trees

Decision trees are a part of classification model but also provides solution to regression problems it is a decision support device which uses tree like structure or a flow chart where each of the inside node depicts the test on set of data values, class labels are depicted by each of the leaf nodes, whereas the branches in a way represents conjunction of features that stretches on way for class labels. Analysis of decision trees are done with algorithmic approach, in this provided dataset splits up into its subsets as per given conditions, basically it is a tree like model that is the manifestation of if-else statements [56]. Tree splits up the source set into its subsets based on each test attribute value. Recursive partitioning is performed on each derived subset that is nothing but repeating the above process on every subset in a recursive manner. the recursion is considered complete when the subset present at the node all possess the same value of the target feature or when no value is added to the forecasted values by splitting. In practice, instances are classified by sorting them form the root to some leaf node down the tree which gives classification of new data, it is classified by initializing from the root node of the tree then examining the variable described by this node and at last going down the tree branch analogous to the value of the instance. Moreover, no domain knowledge or parameter setting is required for the construction of decision tree classifier. Consequently, it becomes appropriate for exploratory knowledge discovery. Generally, decision tree classifiers have good accuracy scores because of their ability of handling high dimensional data with ease. Decision trees are great and very useful in variety of domains. They have emerged as backbone of most of the best performing algorithms in industry [57]. We generally know that they work in stepwise form and have a tree model where, based on some criterion we split up a node but in actual it uses two most important splitting criteria in its background. So, let's understand how they work.

Firstly, talking about Gini impurity, it is the probability of incorrectly categorizing the randomly chosen variable from the set of data. The Gini impurity can be calculated as

$$G = \sum_{i=1}^{c} p(i) \times (1 - p(i)) \tag{9}$$

In above 'c' represents number of classes, 'p(i)' is the probability erratically nominating an element of class 'i'. The quality of split in Gini is determined by weighing the impurities in each branch by how many elements it has. By this, one can conclude higher the Gini gain, better would be the split. The best split is picked by maximizing the Gini gain. Secondly, lets understand the logics of Entropy, it is yet another very famous way of splitting nodes in decision tree. Entropy is simply the measure of randomness in a system it simply tells that how messy is one's data. Entropy basically controls how the decision tree algorithm decides to split up data, it mainly effects how decision tree draws its parameters. Formula for entropy is given as

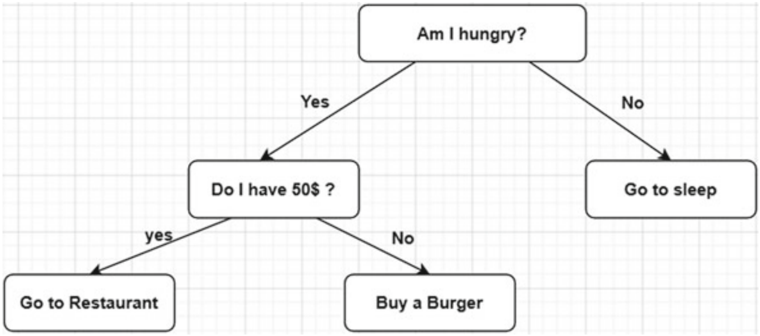


Fig. 4 Simple example of decision tree

$$Entropy = \sum_{i=1}^{c} -p_i \times \log(p_i) \tag{10}$$

where, 'c' is the number of classes in the data and 'pi' are the ratios of variables of each label in the dataset. By using entropy decision tress tidy the data more than they classify them. Yet another important terminology of decision tree is information gain. It can be assessed for a split by subtracting the weighted entropies of every branch from the original entropy. While training a Decision Tree algorithm with the help of these metrics, the best split is nominated by maximizing Information Gain. Figure 4 shows the Simple Example of Decision Tree.

Now, let's comprehend the above with a small example, suppose one person is hungry and wants to make up a decision about what one shall do about the situation then by using decision tree approach the above problem can be solved as.

If the person is not hungry it may go to sleep, and if hungry can have food. Further if person does not possess enough money it may have to buy a burger or else can go to a restaurant. After analyzing the above algorithms and comprehending their working principals, each algorithm individually was implemented for classifying the crop that must be grown in a particular field for maximizing production rate. All the classifier models were compared and best suited classification algorithm that proffers maximum accuracy rates was selected for our further study.

6 Implementation

This section confers about the dataset and various steps which were followed during the implementation process, various data pre-processing steps were performed before model building that were executed using various feature selection and methods, in particular embedded features were applied for extracting the best suited subset amongst all features that were contributing largely for forecasting the target values.

Table 1 First five rows of dataset

Location_Code	Property_Size_Ha	Crop_Group_Name	Crop_Group_Code	Parish_Code
4	0.8	Cereals	600	2
4	0.3	Condiments	400	2
4	9	Condiments	400	2
4	13.25	Condiments	400	2
4	5.06	Condiments	400	2

Embedded features were chosen amongst the other functions because they are much faster, less prone to overfitting and can more accurately define the subsets than other methods discussed in Sect. 3.1. In embedded features itself their various functions to opt from as explained in 3.1.3, but for the model L1 regularization was used, also termed as Lasso regularization, this method makes some of the coefficients of the feature-set zero thereby, implying that zero would be multiplied by some features to calculate the target value.

Now let's look each step of the implementation process in detail starting with data sources, then moving on for data pre-processing and visualizations and finally to results and conclusions. All these stages are mentioned below.

6.1 Gathering Data

The dataset named 'farmer-reports-All-2017-03-20_0917-reportfarmers-parameterized-crop-summary-xlsx' was taken from data.world website [58]. This (data.world) website provides a cutting-edge catalogue for analysing and gathering data. This dataset contains 14 columns and 84,151 rows, there is Crop, aggregated livestock and Property data from around the island. The columns hold feature values, namely 'Location Code', 'Property_Size_Ha', 'Crop_Group_Name', 'Crop_Group_Code', 'Parish_Code', 'Parish_Extension_Code', 'District', 'Parish', 'Parish_Extension_Name', 'Crop_Name', 'Crop_Type_Code', 'Crop_Total_Area_Ha', 'Crop_Count_Total' and 'Farmers'. Data was from the ABIS (Agricultural Business Information System) database having 178,345 overall registered farmers across Jamaica, Country in the Caribbean. Table 1 shows the First five rows of dataset.

6.2 Analyzing Gathered Data

Data Analysis is a crucial part of implementation process, as it aims for checking and visualizing the degree of interdependence of data in process. Data analysis is first and foremost step while implementing any machine learning algorithm this is the procedure of collecting information by use of appropriate application or tool. Which permits you to explore the information and recognize some pattern in it. Based on which one is able to get subtle conclusions, there are various phases of data analysis techniques namely, data requirement collection, data gathering, data cleansing, data analysis, data visualization and data interpretation [59]. In data requirement gathering phase we ought to discover the reason and aim of performing the analysis, here we decide how to examine and what to investigate, we try to understand what type of measures we have to opt for implementing this analysis. After requirements are gathered appropriate data is collected based on the requirements specified. As we collect data from various different sources make sure the data collected is well organized and maintained for processing [60]. Next step is data cleaning or also known as data pre-processing, data collected may possess duplicate records, errors or white space, data must be error free and all white spaces should be replaced with 'null' values before analysis [61]. This step is always done before analysis because with clean and error free data the output of analysis would be closer to expected outcome. After cleaning data, it is being analyzed here we use appropriate tools and software that helps in better understanding of data. Here, we understand, interpret and draw conclusions based on proffered requirement. Another step is data interpretation, data can be interpreted in form of words or maybe as chart or a table these results are used to decide best course of action for further evaluations [62]. Last step in data analysis process is visualization, it normally appears as charts and typical graphs. Provided data is represented graphically so it would be easier and simpler for human brain to apprehend. Certain unknown trends and facts are often revealed by the graphical representation of data. So far, we can conclude data analysis is a means of cleansing, converting and also modelling data for discovering important and useful information for decision making. As discussed earlier this data is raw data and has to be analyzed before processing. For analyzing this data, we followed the above-mentioned steps precisely. Data is being loaded as a csv file format in python script, its empty values and errors are checked, all empty spaces are replaced with 'null' character and degree of interdependence is evaluated. Also, correlation matrix is generated between the 14 given features that tells us what is the degree of relation of every feature with each other. Correlation matrix is a typical table generated that provides correlation coefficients between variables. Each cell tells the value of correlation coefficients between a pair of features. Figure 5 shows the correlation matrix.

After apprehending the correlations amongst the various features, the columns with least contributions in data are dropped with help of python script consequently, making the data cleaner and more meaningful. This cleansed data is used for visualising things more clearly. Our modified data consists of different features namely 'Location Code', 'Property_Size_Ha', 'Crop_Group_Name', 'Crop_Group_Code', 'Parish_Code', 'Parish_Extension_Code', 'District', 'Parish', 'Parish_Extension_Name', 'Crop_Name', 'Crop_Type_Code', 'Crop_Total_Area_Ha' and 'Farmers''. With the help of powerful python tool that is its libraries, using matplotlib from sikit-learn these 3D and 2D graphs are plotted. That are manifesting hidden patterns and trends beside the data.

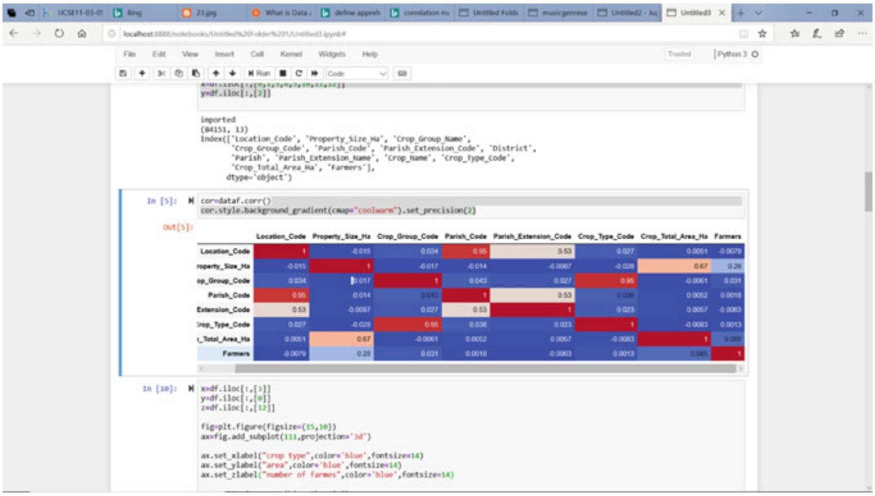


Fig. 5 Correlation matrix

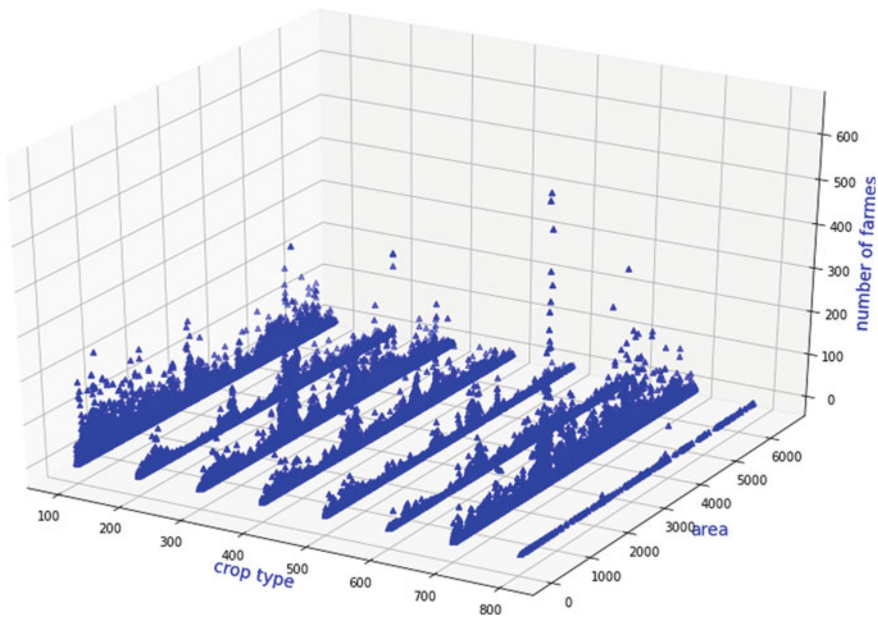


Fig. 6 Plot between crop type, area and number of farmers

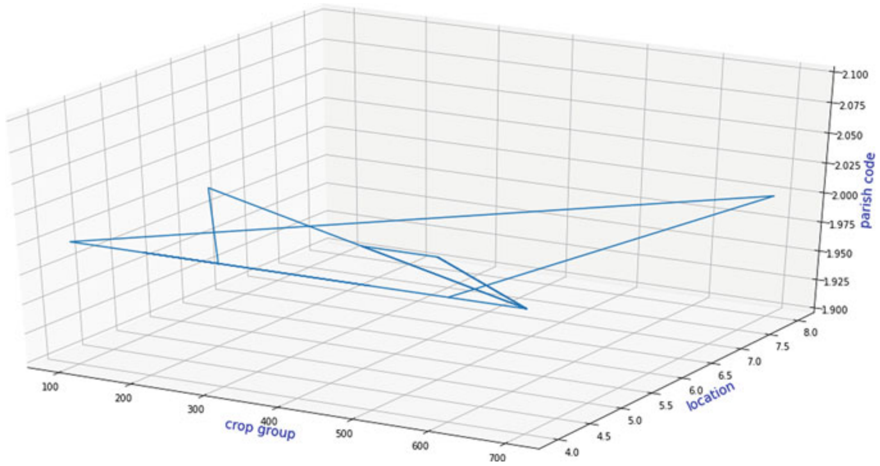


Fig. 7 Graph between crop type, location and parish

Figure 6 is a clear visualization of three different features from the dataset representing co-relations and variations in pattern with respect to one another. Here, features like ‘crop type’, ‘area’ and ‘number of farmers’ are taken to plot a graph in a three-dimensional space.

Figure 7 hereby is a three-dimensional line graph, each of which dimensions are ‘crop group’, ‘location’ and ‘parish code’ respectively. It is representing influence of each of the above fields on one another. Thus, in below graph one dimension is representing group of crops to which it belongs whereas second dimension is representing the place where the field is located and third dimension is denoting the parish code. The farm or holding of land is utilized in fractions or as a whole for agricultural grounds that is located in single parish. The parish of location is that parish where the headquarters are located.

Here, the Fig. 8 is denoting variation of all the numerical feature values present in our data with time. Following graph consists label values on its x-axis and time instances on its y-axis

Figure 9 display pair plot of five most corelated features in dataset. This pair plot represents matrix of bivariate scatter plots between different features. Whereas, the diagonal graphs in the matrix of plots are univariate. The scatter plots help in visualizing the sort of relationship between two features, that is whether they possess positive linear, non-linear, negative linear or no relationship between them.

Figure 10 is a three-dimensional scatter plot, representing influence of three different features namely ‘location’, ‘parish_type’ and ‘number of farmers’. Basically, it depicts variation of one data feature with respect to the data values of other two in three-dimensional space. It is plotted to visualize the data trends and its hidden patterns between these features, it is underlining the fact that which feature is influencing the other and to what extent.

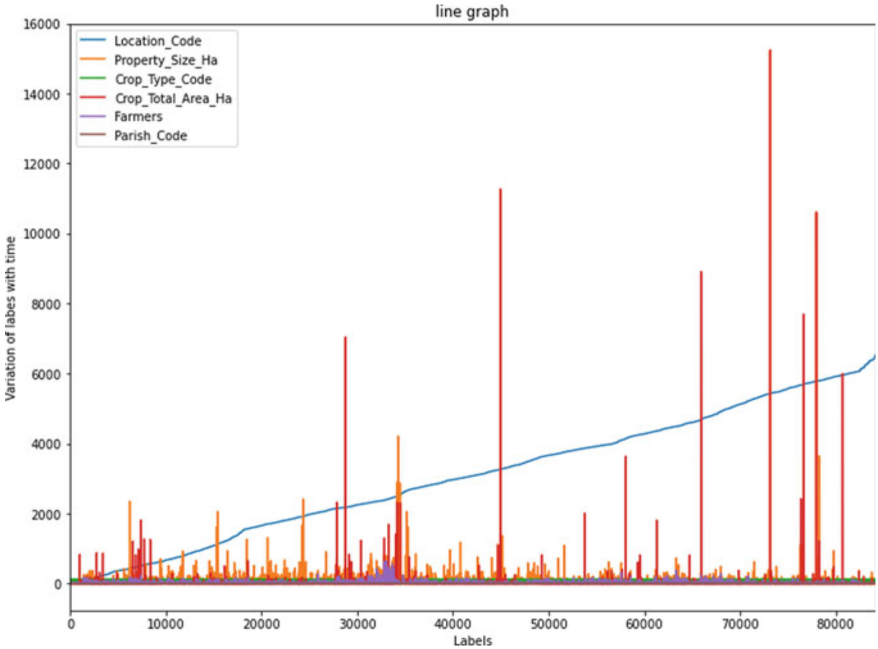


Fig. 8 Data representation between different features

Figure 11 shows a pie chart depicting the amount and types of crops grown. As represented in chart there are eight different groups of crops grown by farmers namely, Tubers, permanent crop, vegetables, cereals, legumes, condiments, fruits and horticulture. Least number of crop group grown is horticulture that is shown with yellow colour whereas, largest group grown is permanent crop shown by turquoise color.

6.3 Results

From the above we interpreted many patterns and relations of different features amongst each other. Now, it's time for predictions and model building. Our major goal of this problem is to classify different types of crops grown on account of other features like area, type of parish, number of farmers working in field etc. this will help us to know what type of crop should be grown in a particular parish to increase the production rates and then the algorithm giving maximum accuracy would be used as classification algorithm for the hybrid approach in further study. To classify these various crops, we have used some powerful classification algorithms that were mentioned above such as k-NN, SVM, logistic regression and decision tree. As every algorithm has different approach works differently with different datasets.

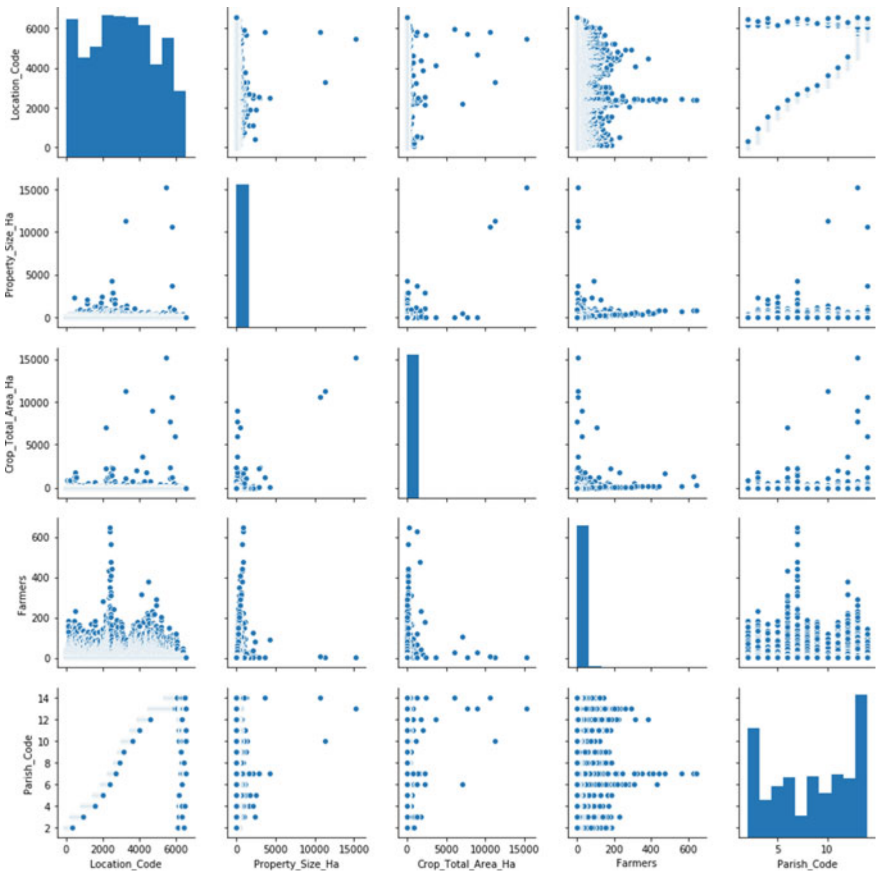


Fig. 9 Display pair plot of five most corelated features in dataset

Starting with k-NN, the algorithm evidenced its high computation capability and fairly powerful prediction potential by producing an accuracy rate of 99.82%. This was mainly due to its principal of close proximity, because of that it happens to create different cluster sets in the data. It assumes that related or similar things in data must lie closer to each other when plotted on a graph. Hence, similar data points will be closer to each other, in k-NN major task is to select the value for variable 'k' which defines the number of cluster formation in the dataset, the value of k was taken to be 8 as we already knew it from the dataset that there are 8 classes of crop therefore, the algorithm foretold all the target values accurately. The next algorithm implemented was logistic regression, this algo predicted the class values with an accuracy rate of 55.66%, which means 45% of forecasted values were fake predictions given by the algorithm. It computes the correlations between the categorial dependent vector and one or more non-dependent vectors using the sigmoid function as explained in Sect. 5.3. This sigmoid function contains values

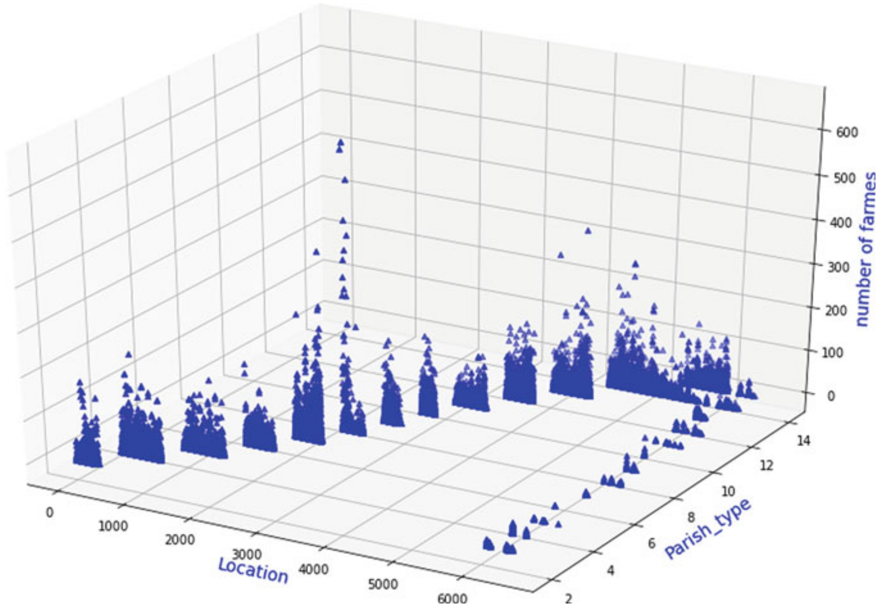


Fig. 10 Plot between location, parish type and number of farmers

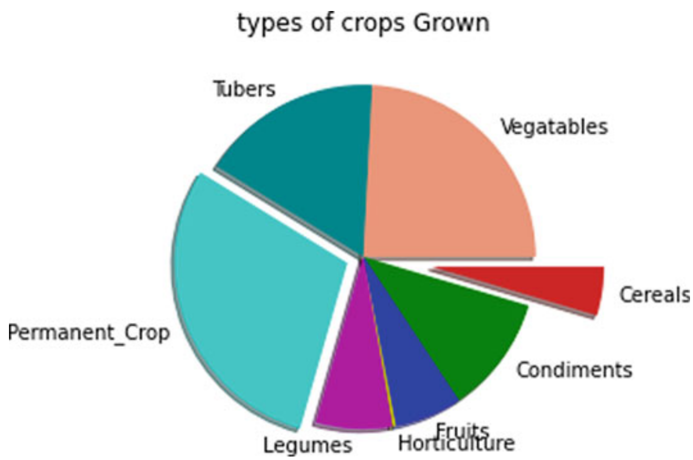


Fig. 11 Different types of crop grown

very close to 0 and 1 across its domain, the weighted sum of inputs evaluates the log of odds or log-odds of a successful event. Moving on further, SVM algo showed a prediction accuracy of 43.36% that was the minimal accuracy amongst all the models. SVM uses iterative approach to generate a hyperplane and classes can be assessed as perpendicular distance from the dividing hyperplane to the support vectors. The

Table 2 Accuracies of different classification algorithms

Model	Accuracy score
k-nearest-neighbour	99.82
Logistic regression	55.66
Support vector machine	43.36
Decision tree	100.0

accuracy of a model solely depends upon the dataset fed into the algorithm, as of SVM, when it comes to high dimensional data as mentioned in Sect. 5.2 selecting the right kernel for the problem is quite difficult also, relatively small set of mislabeled data can considerably decrease model's performance. Finally, decision tree model was executed and predicted the labels with 100.0% accuracy that shows it predicted every value almost correctly. Its principal is to split up dataset into its subsets as per given conditions, basically it is a tree like model that is the manifestation of if-else statements, this breaking down of data into its subsets is done using Gini impurity. Generally, decision tree classifiers have good accuracy scores because of their ability of handling high dimensional data with ease. Best approach in our case comes out to be a decision tree model. It predicted the classes of variables with highest accuracy score. Below are the accuracy scores of various algorithms used. Table 2 shows Accuracies of different classification algorithms.

6.4 *Integrating IoT and ML for Intelligent Farming—A Future Approach*

Suitable irrigation is important aspect for proper growth and development of the crops. Irrigation systems are installed in crop fields so as to maintain the moisture and humidity levels of the land. Typically saying, there exists two types of irrigation systems namely, conventional methods and modern irrigation methods. Irrigating farmers using buckets, pumps, canals etc. goes under conventional methods. Conversely, modern irrigation systems or the current trends for irrigation includes sprinkler systems, pot irrigation systems, drip irrigation etc. As different crops require different water levels to develop. These irrigation systems lack the capability to predict the precise amount of water needed for the particular crop at an instance of time. They solely depend upon the knowledge and experience of the farmer. For improving this drawback smart irrigation systems are developed using internet of things integrated with power of machine learning. These systems possess the power to take decisions and act upon the given conditions. These systems are automated by IoT and ML driven environment. Both IoT and ML can be put together in order to mutually compose a smart irrigation system. Providing right amount of water to the fields will help the farmer to maintain soil sanity and flourish crops that have quality and value. The intelligent system requires both hardware and software tools however, hardware used in composing this system are Arduino, sensors, Bolt IoT kit and server motor. Arduino is a microcontroller which is used to control and operate electronic devices like server motors. Whereas, various types of sensors are used namely, light sensor also known as ‘photo sensor’, rain sensor, temperature sensor and water level sensor. The light sensors convert the radiated light energy belonging to any spectrum of light to an electrical signal, rain sensors would detect rain and its amount and would report back the data to software. Temperature sensors gathers temperature rates of the source in our case it is soil, these sensors would collect soil temperatures and data is reverted back to system. Water levels of soil would reveal its humidity and moisture content at an instance, this data is gathered by water level sensors. The interface for communication between these hardware devices and software environment is simulated by Bolt IoT kit. Collected information from the sensors is then fed into the ML model for prediction. The model that has predicted the crop type would now is used to predict the water level content for that particular crop. In this case we would use decision tree classifier as it had the maximum accuracy. New data containing information about the crop type, moisture content, soil nature, humidity, temperature of soil etc. is provided to the decision tree model for classifying whether the crop field needs to be irrigated or not. This newly generated model is fed with the data coming from the field, in turn the model would predict the soil water levels needed for that particular crop to grow. If these predicted value matches with the current water levels of the field then there is no need to water the crops. If the current water level values are less than the required predicted value of water levels then there is a need to irrigate the field till the value of moisture level reaches the required value that has been predicted by the decision tree model. In this way, IoT was integrated with ML to construct intelligent irrigation system.

7 Conclusion and Future Scope

This chapter discussed some of the important data analysis techniques that we enacted for estimating crop types. Also, we deliberated upon erection of IoT and ML based system that can help farmers to cultivate crops more efficiently and with boosted production. Our model (decision tree) came out to be greatly effective while predicting class labels, also we came across some of the most important mathematical evaluations that these algorithm uses for computation of forecasting values, as all algorithms have different approaches to solve given problem they behave differently with different data, that means the choice of model in use solely is determined by data in use. These algorithms perform distinctly with data having different number of dimensions and different correlations amongst its features. In our case k-NN and decision trees put forward best case scenarios for the proffered problem. Also, with the help of sensors and Arduino microcontroller we were able to formulate an intelligent irrigation system which checks the soil’s current water level and moisture contents further with the help of ML model it can compute whether the crop field is in need for irrigation. This study can be extended for more information and evaluations by using neural networks, deep learning techniques and computer vision.

References

1. Li, S., Peng, S., Chen, W., Lu, X.: INCOME: Practical land monitoring in precision agriculture with sensor networks. Comput. Commun. **36**(4), 459–467 (2013). https://doi.org/10.1016/j.comcom.2012.10.011
2. Jones, A.D., Ngure, F.M., Pelto, G., Young, S.L.: What are we assessing when we measure food security? A compendium and review of current metrics. Adv. Nutr. **4**(5), 481–505 (2013). https://doi.org/10.3945/an.113.004119
3. Gm, H., Gourisaria, M.K., Pandey, M., Rautaray, S.S.: A comprehensive survey and analysis of generative models in machine learning. Comput. Sci. Rev. **38**, (2020). https://doi.org/10.1016/j.cosrev.2020.100285
4. Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V., Kumar, A.: A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Comput. Oper. Res. **119**, (2020). https://doi.org/10.1016/j.cor.2020.104926
5. Klompenburg, T.V., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agric. **177**, (2020). https://doi.org/10.1016/j.compag.2020.105709
6. Elavarasan, D., Vincent, D.R., Sharma, V., Zomaya, A.Y., Srinivasan, K.: Forecasting yield by integrating agrarian factors and machine learning models: a survey. Comput. Electron. Agric. **155**, 257–282 (2018). https://doi.org/10.1016/j.compag.2018.10.024
7. Mohanta, B.K., Jena, D., Satapathy, U., Patnaik, S.: Survey on IoT security: challenges and solution using machine learning, artificial intelligence and blockchain technology. Internet Things **11**, (2020). https://doi.org/10.1016/j.iot.2020.100227
8. Ogutu, G.E., Franssen, W.H., Supit, I., Omondi, P., Hutjes, R.W.: Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts. Agric. For. Meteorol. **250–251**, 243–261 (2018). https://doi.org/10.1016/j.agrformet.2017.12.256
9. Holzman, M.E., Carmona, F., Rivas, R., Niclòs, R.: Early assessment of crop yield from remotely sensed water stress and solar radiation data. ISPRS J. Photogram. Remote Sens. **145**, 297–308 (2018). https://doi.org/10.1016/j.isprsjprs.2018.03.014
10. Singh, A., Ganapathysubramanian, B., Singh, A.K., Sarkar, S.: Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci. **21**(2), 110–124 (2016). https://doi.org/10.1016/j.tplants.2015.10.015
11. Xie, J., Yu, F.R., Huang, T., Xie, R., Liu, J., Wang, C., Liu, Y.: A survey of machine learning techniques applied to software defined networking (SDN): research issues and challenges. IEEE Commun. Surv. Tutor. **21**(1), 393–430 (2019). https://doi.org/10.1109/comst.2018.2866942
12. Whetton, R., Zhao, Y., Shaddad, S., Mouazen, A.M.: Nonlinear parametric modelling to study how soil properties affect crop yields and NDVI. Comput. Electron. Agric. **138**, 127–136 (2017). https://doi.org/10.1016/j.compag.2017.04.016
13. Al-Garadi, M.A., Mohamed, A., Al-Ali, A.K., Du, X., Ali, I., Guizani, M.: A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Commun. Surv. Tutor. **22**(3), 1646–1685 (2020). https://doi.org/10.1109/comst.2020.2988293
14. Dash, Y., Mishra, S.K., Panigrahi, B.K.: Rainfall prediction for the Kerala state of India using artificial intelligence approaches. Comput. Electr. Eng. **70**, 66–73 (2018). https://doi.org/10.1016/j.compeleceng.2018.06.004
15. Wieder, W., Shoop, S., Barna, L., Franz, T., Finkenbiner, C.: Comparison of soil strength measurements of agricultural soils in Nebraska. J. Terrramech. **77**, 31–48 (2018). https://doi.org/10.1016/j.jterra.2018.02.003
16. Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., Li, Z.: A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sens. Environ. **210**, 35–47 (2018). https://doi.org/10.1016/j.rse.2018.02.045

17. Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., Mouazen, A.: Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. **121**, 57–65 (2016). https://doi.org/10.1016/j.compag.2015.11.018
18. Rehman, T.U., Mahmud, M.S., Chang, Y.K., Jin, J., Shin, J.: Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. Comput. Electron. Agric. **156**, 585–605 (2019). https://doi.org/10.1016/j.compag.2018.12.006
19. Dhar, S., Cherkassky, V.: Development and evaluation of cost-sensitive universum-SVM. IEEE Trans. Cybern. **45**(4), 806–818 (2015). https://doi.org/10.1109/tcyb.2014.2336876
20. Khan, M.M., Arif, R.B., Siddique, M.A., Oishe, M.R.: Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from uci machine learning repository. In: 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT) (2018). https://doi.org/10.1109/ceeict.2018.8628041
21. Ravi Shankar, N., Jaya Pramanik, S.C., Kumar, J., Singh, D.R., Bibi, N., Nawaz, S., Biswas, T.K.: Study on integrated farming system (IFS) under different resource condition of Island Ecosystem. J. Farming Syst. Res. Develop. **13**(1), 1–9 (2007)
22. Botsiou, M., Dagdilelis, V.: Could ICT contribute to the revival of family farming? Int. J. Agric. Resour. Govern. Ecol. **10**(2), 112 (2014). https://doi.org/10.1504/ijarge.2014.063579
23. Song, X., Fan, G., Rao, M.: Automatic CRP mapping using nonparametric machine learning approaches. IEEE Trans. Geosci. Remote Sens. **43**(4), 888–897 (2005). https://doi.org/10.1109/tgrs.2005.844031
24. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Front. Plant Sci. (2016). https://doi.org/10.3389/fpls.2016.01419
25. Kumar, M., Nagar, M.: Big data analytics in agriculture and distribution channel. In: 2017 International Conference on Computing Methodologies and Communication (ICCMC) (2017). https://doi.org/10.1109/iccmc.2017.8282714
26. Huang, J., Zhang, L.: The big data processing platform for intelligent agriculture (2017). 10.1063/1.4992850
27. Varman, S.A., Baskaran, A.R., Aravindh, S., Prabhu, E.: Deep learning and IoT for smart agriculture using WSN. In: 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (2017). https://doi.org/10.1109/iccic.2017.8524140
28. Payero, J.O., Mirzakhani-Nafchi, A., Khalilian, A., Qiao, X., Davis, R.: Development of a low-cost Internet-of-Things (IoT) system for monitoring soil water potential using watermark 200SS Sensors. Adv. Internet Things **07**(03), 71–86 (2017). https://doi.org/10.4236/ait.2017.73005
29. Byakatonda, J., Parida, B., Kenabatho, P.K., Moalafhi, D.: Influence of climate variability and length of rainy season on crop yields in semiarid Botswana. Agric. For. Meteorol. **248**, 130–144 (2018). https://doi.org/10.1016/j.agrformet.2017.09.016
30. Nevavuori, P., Narra, N., Lipping, T.: Crop yield prediction with deep convolutional neural networks. Comput. Electron. Agric. **163**, 104859 (2019). https://doi.org/10.1016/j.compag.2019.104859
31. Yang, Q., Shi, L., Han, J., Zha, Y., Zhu, P.: Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. Field Crops Res. **235**, 142–153 (2019). https://doi.org/10.1016/j.fcr.2019.02.022
32. Cynthia, S.T., Hossain, K.M., Hasan, M.N., Asaduzzaman, M., Das, A.K.: Automated detection of plant diseases using image processing and faster R-CNN algorithm. In: 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI) (2019). https://doi.org/10.1109/sti47673.2019.9068092
33. Samsonovich, A.V.: Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. Cogn. Syst. Res. **60**, 57–76 (2020). https://doi.org/10.1016/j.cogsys.2019.12.002
34. Hang, X., Gao, H., Jia, S.: Identification of tomato diseases using skip-gram and LSTM based on QA(Question-Answer) system. J. Phys: Conf. Ser. **1437**, 012048 (2020). https://doi.org/10.1088/1742-6596/1437/1/012048

35. Shafi, U., Mumtaz, R., Iqbal, N., Zaidi, S.M., Zaidi, S.A., Hussain, I., Mahmood, Z.: A multimodal approach for crop health mapping using low altitude remote sensing, internet of things (IoT) and machine learning. IEEE Access **8**, 112708–112724 (2020). https://doi.org/10.1109/access.2020.3002948
36. Elavarasan, D., Vincent, P.M.: Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access **8**, 86886–86901 (2020). https://doi.org/10.1109/access.2020.2992480
37. Anwar, M.M., Farooqi, S., Qureshi, Y.: Agriculture sector performance: an analysis through the role of agriculture Sector share in GDP. J. Agricult. Econ., Extension Rural Develop. **3**(3), 270–275 (2015)
38. Shafi, U., Mumtaz, R., García-Nieto, J., Hassan, S.A., Zaidi, S.A., Iqbal, N.: Precision agriculture techniques and practices: from considerations to applications. Sensors **19**(17), 3796 (2019). https://doi.org/10.3390/s19173796
39. Wachowiak, M.P., Walters, D.F., Kovacs, J.M., Wachowiak-Smolíková, R., James, A.L.: Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. Comput. Electron. Agric. **143**, 149–164 (2017). https://doi.org/10.1016/j.compag.2017.09.035
40. Wrapper-Mediator Approach to Query Evaluation. (n.d.). SpringerReference. https://doi.org/10.1007/springerreference_63161
41. Shahzadi, R., Ferzund, J., Tausif, M., Asif, M.: Internet of things based expert system for smart agriculture. Int. J. Adv. Comput. Sci. Appl. **7**(9), (2016). https://doi.org/10.14569/ijacsa.2016.070947
42. Roopaei, M., Rad, P., Choo, K.R.: Cloud of things in smart agriculture: intelligent irrigation monitoring by thermal imaging. IEEE Cloud Comput. **4**(1), 10–15 (2017). https://doi.org/10.1109/mcc.2017.5
43. Mchugh, M.L.: The Chi-square test of independence. Biochemia Medica (2013). https://doi.org/10.11613/bm.2013.018
44. Chlingaryan, A., Sukkarieh, S., Whelan, B.: Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. **151**, 61–69 (2018). https://doi.org/10.1016/j.compag.2018.05.012
45. Satorra, A., Bentler, P.M.: A scaled difference chi-square test statistic for moment structure analysis. Psychometrika **66**(4), 507–514 (2001). https://doi.org/10.1007/bf02296192
46. Chen, Y., Chen, M.C.: Using chi-square statistics to measure similarities for text categorization. Expert Syst. Appl. **38**(4), 3085–3090 (2011). https://doi.org/10.1016/j.eswa.2010.08.100
47. Patel, S.P., Upadhyay, S.: Euclidean distance-based feature ranking and subset selection for bearing fault diagnosis. Expert Syst. Appl. **154**, 113400 (2020). https://doi.org/10.1016/j.eswa.2020.113400
48. Bah, M.D., Dericquebourg, E., Hafiane, A., Canals, R.: Deep learning based classification system for identifying weeds using high-resolution UAV imagery. Adv. Intell. Syst. Comput. Intell. Comput. (2018). https://doi.org/10.1007/978-3-030-01177-2_13
49. Jiang, Z., Huete, A.R., Chen, J., Chen, Y., Li, J., Yan, G., Zhang, X.: Analysis of NDVI and scaled difference vegetation index retrievals of vegetation fraction. Remote Sens. Environ. **101**(3), 366–378 (2006). https://doi.org/10.1016/j.rse.2006.01.003
50. Ahmed, S., Lee, Y., Hyun, S., Koo, I.: Covert cyber assault detection in smart grid networks utilizing feature selection and euclidean distance-based machine learning. Appl. Sci. **8**(5), 772 (2018). https://doi.org/10.3390/app8050772
51. Mesquita, D.P., Gomes, J.P., Junior, A.H., Nobre, J.S.: Euclidean distance estimation in incomplete datasets. Neurocomputing **248**, 11–18 (2017). https://doi.org/10.1016/j.neucom.2016.12.081
52. Popescu, D., Stoican, F., Stamatescu, G., Ichim, L., Dragana, C.: Advanced UAV–WSN system for intelligent monitoring in precision agriculture. Sensors **20**(3), 817 (2020). https://doi.org/10.3390/s20030817

53. Harjani, M., Grover, M., Sharma, N., Kaushik, I.: Analysis of various machine learning algorithm for cardiac pulse prediction. In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (2019). https://doi.org/10.1109/icccis48478.2019.8974519
54. Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F.: Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric. For. Meteorol. **218–219**, 74–84 (2016). https://doi.org/10.1016/j.agrformet.2015.11.003
55. Kaya, A., Keceli, A.S., Catal, C., Yalic, H.Y., Temucin, H., Tekinerdogan, B.: Analysis of transfer learning for deep neural network-based plant classification models. Comput. Electron. Agric. **158**, 20–29 (2019). https://doi.org/10.1016/j.compag.2019.01.041
56. Kamilaris, A., Prenafeta-Boldú, F.X.: Deep learning in agriculture: a survey. Comput. Electron. Agric. **147**, 70–90 (2018). https://doi.org/10.1016/j.compag.2018.02.016
57. Rustagi, A., Manchanda, C., Sharma, N., Kaushik, I.: Depression anatomy using combinational deep neural network. Adv. Intell. Syst. Comput. Int. Conf. Innov. Comput. Commun. (2020). https://doi.org/10.1007/978-981-15-5148-2_3
58. https://data.world/jamaica/farmer-reports
59. Usama, M., Ahmad, B., Yang, J., Qamar, S., Ahmad, P., Zhang, Y., Guna, J.: REMOVED: equipping recurrent neural network with CNN-style attention mechanisms for sentiment analysis of network reviews. Comput. Commun. **148**, 98 (2019). https://doi.org/10.1016/j.comcom.2019.08.002
60. Gheisari, S., Tahavori, E.: CCCLA: a cognitive approach for congestion control in Internet of Things using a game of learning automata. Comput. Commun. **147**, 40–49 (2019). https://doi.org/10.1016/j.comcom.2019.08.017
61. Sharma, A., Sharma, N., Kaushik, I., Kumar, S., Khatoon, N.: Predictive analysis of type 2 diabetes using hybrid ML model and IoT. In: IoT Security Paradigms and Applications (pp. 303–320). CRC Press
62. Kathuria, R.S., Gautam, S., Singh, A., Khatri, S., Yadav, N.: Real time sentiment analysis on twitter data using deep learning (Keras). In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India (2019), pp. 69–73, https://doi.org/10.1109/icccis48478.2019.8974557

Soft Computing Techniques in Cloud Computing and Computer Networking

Hybrid Cloud Data Protection Using Machine Learning Approach

D. Praveena, S. Thanga Ramya, V. P. Gladis Pushparathi, Pratap Bethi, and S. Poopandian

Abstract In today's digital world, information created by Internet of Things (IoT) devices has expanded drastically. This expansion is due to an increase in the number of IoT devices associated with the internet. Hybrid cloud computing provides enormous support to these emerging IoT devices in processing vast data. However, security is a challenging issue because of the integration of IoT and hybrid cloud. To achieve a sufficient level of hybrid cloud IoT security, a combination of Enhanced C4.5 machine learning algorithm and Dynamic Spatio Role-Based Access Control Algorithm is introduced. In this approach, the data users are classified using the Enhanced C4.5 algorithm and the user's level of cloud data access is restricted using the Dynamic Spatio Role-Based Access Control Algorithm. As a result, the major security issues pertaining to IoT cloud are addressed. The security framework also uses a deduplication algorithm for eliminating redundant data and significantly improving IoT data storage requirements.

Keywords Internet of Things (IoT) · Machine learning · Hybrid cloud networks · Security · Deduplication · Access control · Secured storage

D. Praveena (✉) · S. Thanga Ramya
Department of Information Technology, R.M.D Engineering College, Kavaraipettai 601206, India
e-mail: praveena.it@rmd.ac.in

S. Thanga Ramya
e-mail: str.it@rmd.ac.in

V. P. Gladis Pushparathi
Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai, India

P. Bethi
OPG Power Generation Pvt. Ltd., Chennai, India

S. Poopandian
Infosys Pvt., Ltd., Chennai, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_7

1 Introduction

The major concern for hybrid cloud computing is security challenges despite the advantages of on demand self-service, pay per use, vast network access and resource elasticity [1]. To provide security to the data, existing mechanism encrypts the data before the data is stored in cloud service provider. However, the same data can be duplicated by different users with different encrypted values at the cloud. In spite of the vast cloud storage space, duplication waste resources, power and increases the complexity of management of data. Thus deduplication has become necessary task for efficient cloud storage. Deduplication effectively save cost and space up to 90–95%, thereby reducing the burden of the cloud users.

Deduplication over the encrypted data is provided by convergent encryption. In convergent encryption (CE), the encryption key is obtained by hashing the input file. Then the encryption key is used to encrypt data and gets cipher text for cloud server storage. Similar cipher text is yielded for the same file hash value; thus deduplication is performed. However, convergent encryption is subject to offline brute-force dictionary attacks [2]. DupLess technique avoids brute force attack and ensures high security as long as the key server which aids in data encryption is secured from the attacks.

Convergent encryption security threats can also be solved by Cloud Dedup [3] but still it cannot handle the issues of data deletion. An information holder that expels the information from the cloud can even now get to similar information since it despite everything knows the information encryption key if the information isn't totally expelled from the cloud. The existing solutions the reduces the storage overhead to greater extent but still data users find in difficulty access control in data storage in a convenient way.

2 Problem Statement

Cloud security is still a hard issue in recent days particularly for the hybrid cloud because of the mixture of private and public cloud. Many existing security mechanisms available are heavy time consuming, privacy issues and lacks flexible access control [4]. To solve the major hybrid cloud issues, a new approach is proposed which uses a machine learning technique for efficient data storage and retrieval. This approach uses newly proposed deduplication algorithm with Dynamic Spatial Role Based Access Control and enhanced C4.5 machine learning algorithm [5] for restricting unauthorized user accessing the cloud without redundancy. The user can download the requested data in a flexible manner, once the user access rights are verified.

3 De-duplication

Deduplication, also called single instance technique, is a compression technique for expelling duplicate data and storing the original data copy with the goal that it will save the storage space to protect sensitive data [6]. The information security and access to specific data are particularly vital in recent times. Therefore, the features in deduplication have been generally used in the cloud storage system [2]. The management of the consistently expanding volume of data is critically tricky for cloud storage system; the data management makes conceivable through deduplication concept. In the organizations, the storage systems of several part contain multiple copies of same data [7]. The different users save the same file might at various places, which means multiple copies are getting created which has the similar data. Deduplication dispenses these multiple copies which get created by storing a data copy and a pointer pointing to the other copies of the actual file. Deduplication technique is utilized in organizations for disaster recovery and backup applications; this strategy also used to make free space in the storage systems [6].

Deduplication lessens the needs of storage by up to 90–95% for the application of reinforcement, 68% in standard file framework and therefore we generate more space for our file [8]. Essential problems in data deduplication are to give privacy and security to shield the data from insider or outsider attacker. Figure 1 shows the deduplication process before accumulating data into the memory.

The two major types of deduplication, depending on granularity are block-level deduplication and deduplication based on file-level. The entire file is considered for file-level deduplication, thereby a small change or update creates a new file which is different from older version of it, resulting in reduction of duplication ratio. But data blocks are considered for deduplication in block-level deduplication.

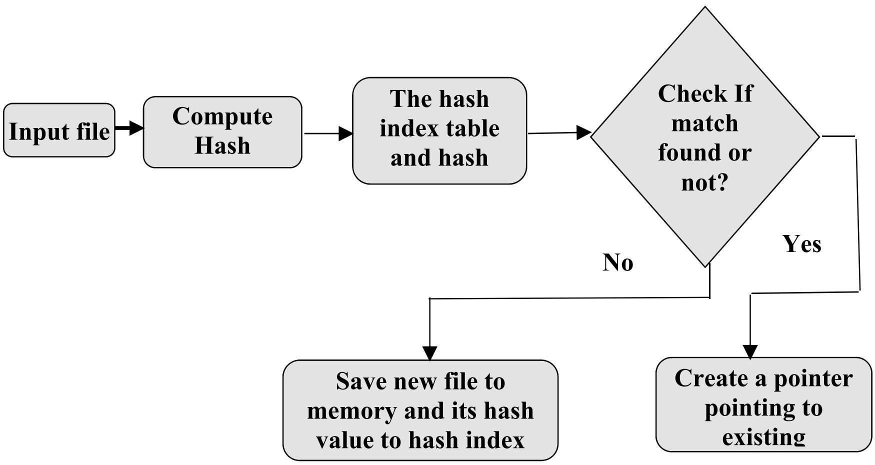


Fig. 1 Data deduplication flowchart

Location of deduplication can also be considered for dividing the deduplication, i.e., as source-side deduplication and as client-side deduplication. Bandwidth is guaranty saved in client-side deduplication guarantees as only file's hash value is transferred to the server if the duplicate exists [9].

Deduplication is most utilized in different applications like primary storage, backup, metadata management and so on for storage optimization [10]. Data deduplication has turned into an essential part of controlling repositories of external data to the cloud data centres. In any case, centralized data centres face issues of data loss and accessibility if something goes flawed as deduplication keeps up only a unique copy of the content.

When a user needs to download a document from the cloud, he can only login into the system and select file for download that is in encrypted file form, so user decrypts this file utilizing convergent key. To make information administration flexible in distributed computing, deduplication has been an understood system and has become much thoughtful at recent times [8]. Information deduplication is a technique applied on certain information and pressure procedure for removing copy duplicates of rehashing information. The procedure is employed to improve stockpiling us-age and can similarly be associated with network information transferring to decrease the quantity of bytes that must be sent. Deduplication eliminates bounty information, by putting away one duplicate of physical information and diverting other recreate information to that physical duplicate without keeping different duplicates of same information [11].

4 Cloud Security Protection Framework with Machine Learning Modules

The overall framework of the proposed method is shown in Fig. 2. The security system has four diverse subcomponents to be specific access control stage, de-duplication stage, area guide chief. The information assortment module has gathered the cloud client information and the association's reports and put away in the hybrid cloud storage. The duplicate data is expelled, and the scrambled de-duplication module encodes the accessible information that are as of now put away in the hybrid cloud.

The hybrid cloud data recovery is the responsibility of encrypted reduplication, that are as of now accessible in the cloud storage and evacuating the duplication data and furthermore in encrypting the recovered information.

The UI block is responsible for storing enciphered data in the hybrid cloud storage without replication. From the rundown of security dangers, redundancy is one of the noteworthy dangers. The cloud user stores the organization data in the hybrid cloud database via the user interface module with approved authorization control of access control. The cloud users are restricted in accessing the cloud storage by efficient access control module. The module permits only the authorized user of the organization with approved credentials.

Access control algorithm and location manager are the two components of the access restriction control system in the security framework. The position administrator shows the cloud clients area wherefrom they are getting to the information by different cloud clients. Here, the UI module acts as an interface and is in charge for communicating with all the modules of security framework, for example, cloud storage, encrypted de-duplication module, and components of access control management.

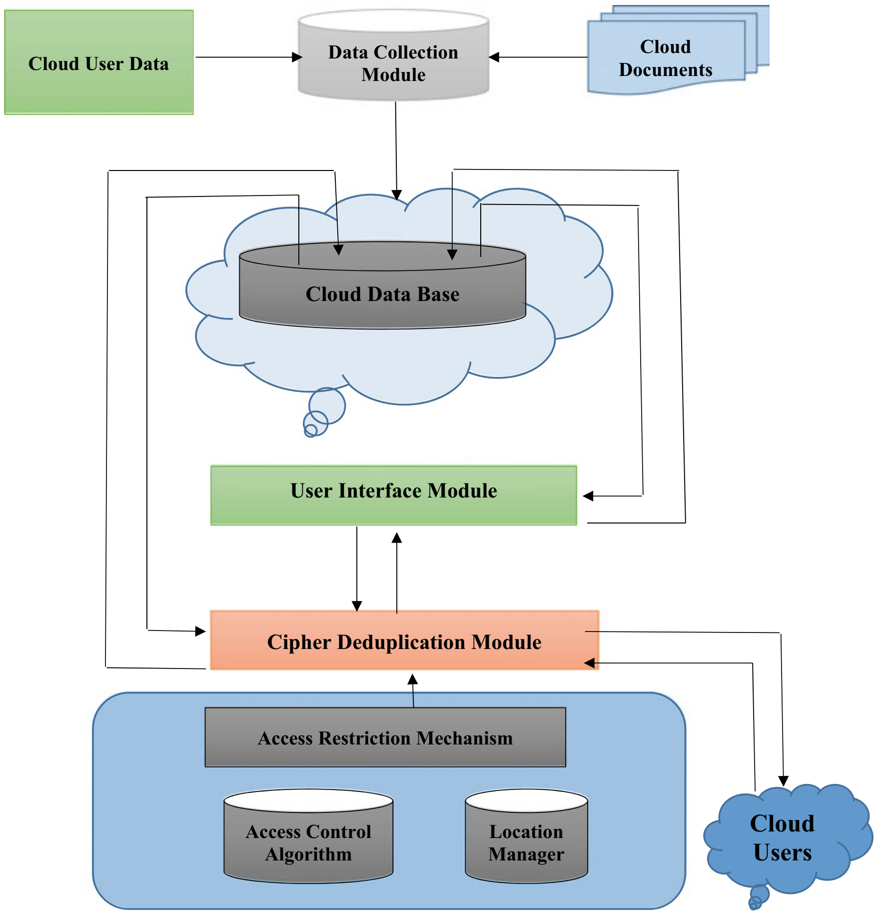


Fig. 2 Hybrid cloud data protection with a machine learning application

4.1 Cloud Client Classification Using Enhanced C4.5 Algorithm

Maintaining a security management of person authorizations is an essential element of system. In a multi-area or a dynamic environment, those essential constructing modules of get entry access to manage aren't so well hooked up. RBAC nevertheless lacks the capacity to address evolving context element and managing any unexpected situations [12]. For example, whilst a brand new person requests a get right of entry to and sources they want may also vary, conventional get right of entry to manage mechanisms lacks capacity to address those form of requests effectively. In this situation, it has to now no longer be vital to absolutely extrude the hooked up local get right of entry to manage mechanisms for a new client. Instead, to aid dynamic environments, it has to be viable to automate the era and enforcement of get right of entry to manage guidelines to satisfy the new user needs, even as the existing mechanism continue to be intact.

Normally, when another client is added to a current framework, manual intercession of security director gets important to choose whether a client should access a mentioned asset or not. In a role based access control security condition, if the user's access rights are still unknown or the mentioned access' point of view are not significant, it gets hard for the security overseer to decide a reasonable job for another client through which a client can get to benefits that are required to play out a specific activity [13]. From security point of view, it is critical to guarantee that conceded access to another client is steady with the hierarchical access control strategy. Be that as it may, a security overseer can commit an error in allotting proper job to another client. In this manner, when a RBAC is set up, deciding a reasonable job for another client is a basic yet testing task. As shown in Fig. 3, the UI module has a current AI technique known as Improved C4.5.

The UI module has a current AI technique known as Improved C4.5 scheme [14] to characterize the cloud users who request information from the cloud database. The model grouping of a user for any role is developed by utilizing the data about permission of existing clients who have that role, and furthermore the qualifications of existing clients who don't have that role. As it were, labelled training information (supervised learning) for roles are first utilized for model structure process and later those models are utilized to characterize a client for a possible role.

A few classification techniques have been proposed in the writing of data mining and AI. Decision tree is ordinarily utilized as a choice help apparatus that utilizes a tree-like chart or model of choices and conceivable result of those choices for the choice investigation reason. Improved C4.5 calculations depends on isolate and overcome technique where issue of taking in arrangement structure from known arrangement of perception prompts a methodical development of a decision tree.

The existing system involves three steps to add a new user to the overall process:

1. For each role, build a classification model.
2. Selecting roles having requested permission.
3. For each role external user's attributes are classified.

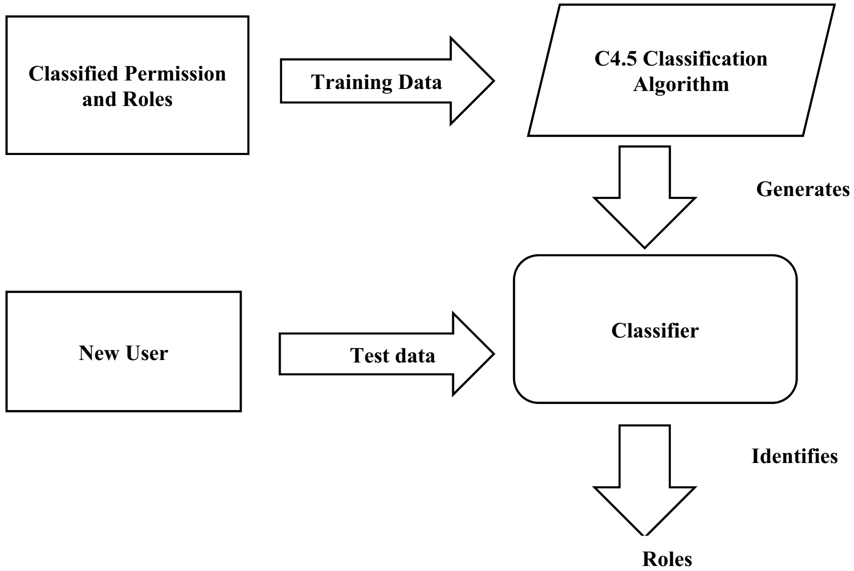


Fig. 3 User role classification process using C4.5 algorithm

C 4.5 Algorithm for assigning role to a new user

Input: PA: Permission role assignment

Output: Assignment of role to new user_role

Step 1: new user_role→ empty

Step 2: for each role in PA do

    Step 2.1: If the role has permission on a file then

    Step 2.2: new user_role→ role

Step 5: End

4.2 De-duplication Processing Algorithm

The novel de-duplication algorithm removes the replicated contents that are existing in the hybrid cloud. This enables access in matching and eliminating the similar files that are available in the hybrid cloud storage. The cryptographic features in the deduplication processing algorithm are utilized for the encryption of the data.

Encrypted Data Upload: Data owner encrypts the file to be uploaded using randomly selected symmetric key using AES-256 algorithm and stores only the encrypted file in the hybrid cloud storage. Further for each file the CSP computes the hash token with SHA-512 for the file and stores hash token along with encrypted symmetric key as record in cloud storage [15]. Later the data owner can share the encrypted symmetric key for the valid data user using RSA public key cryptography.

Data De-duplication: When the same files are to be stored in the cloud storage, the data deduplication occurs. CSP compares the hash token for the new file with the records in the cloud data [16]. If the file already exists, the CSP checks the access rights of the data holder and if valid permits the symmetric key to decrypt the encrypted data.

Data Updation: The data owner updates encryption key to all the data user, whenever the new update encryption is updated.

Data Deletion: When CSP receives the data deletion request from user A, the CSP checks the eligibility, then deletes the record of A and its access rights to the file. Further it checks that the deduplication records are empty, if it is empty, it removes the encrypted data and all its records.

4.3 Dynamic Spatial RBAC Algorithm

Dynamic Spatial RBAC Algorithm provides to restricted and secure access to the cloud users for accessing the hybrid cloud files. DSRBACA is capable of assigning a set of roles described by set of permission to different cloud users dynamically under various constraints. Permissions are approval to execute set of operation on one or more data and depends on the role and owner location. This role is periodically changed the location of the cloud user is identified and monitored dynamically for fixing their rights to access their data organization.

The user permissions are validated with the access rights and in addition, the spatio constraints with the status level of the user are used for validating spatio action status access control policies [17]. Status level is a named position of a user with other users in a specific domain. For example, a bank manager and clerk has two distinct status levels in bank domain where the manager status level is higher than clerk in accessing files. In DSRBACA, a user's status level is determined from the ascribed status and action status at a particular location. The assignment of user status level and access privileges changes dynamically with respect to the change in the location.

Algorithm for Deduplication Process

First Phase: Data Upload

Step 1: Calculate the hash token for the user input file using SHA-512 algorithm

Step 2: Apply symmetric encryption algorithm AES-256 to encrypt the file with key k

Step 3: Apply RSA algorithm to encrypt the symmetric key K used in above step

Step 4: Store the encrypted input file into the database along with hash token and encrypted key.

Step 5: End

Second Phase: Deduplication

Step 1: Calculate the hash token for the files from the database

Step 2: Check whether hash token already exist in the database

Step 3: It checks whether a different user is storing the same data, if yes, the CSP performs deduplication,

Step 4: If no, hash token of the encrypted content along with the key will be stored in the cloud database.

Step 5: End

The steps involved in Dynamic Spatial RBAC Algorithm are as follows:

Dynamic Spatial RBAC Algorithm

1: Read→cloud users requirement

2: user requirements are initialized in queue

3: For accessing the data, Read →cloud users requirement

4: While (! isEmpty(QUEUE))

do

5: From the queue find the next user

6: Verify→ cloud user query by using the database access privileges and the user level

7: Check whether (user level >5), then

8: Verify→ access rights→use time and location to assign the suitable role

9: Update user level based on the time and location

10: If (check the time and spatial constraints to validate cloud user rights)

11: Check if (valid cloud user), then

    Find the address of the data stored and its position in the hash table

    Input→ respective position with content and key

    Decrypt→the stored encrypted data with RSA

    Show→ original message

Else

    Show→ ERROR.

12: Repeat → until 4 to 11 until the no user in queue.

13: End

Dynamic Spatial RBAC Algorithm utilizes RSA data encryption and the cloud datasets for time and spatial requirements is assessed by the implementation for spatio access control mechanism.

5 Experimental Results

An OpenStack cloud is used to implement the security system framework and quantities of trials are assessed. In implementation, the trials were led utilizing a 2.4 GHz Intel I5 processor running at, 7200 RPM Western Digital, 4 GB RAM, and 8 MB cradle 350 GB Serial ATA drive. It includes, information examining process was completed to assess the exhibition of the framework. In the results five different

Table 1 Analysis of users request denial

Sl. No.	Count of tried user request	Count of requests denied by	
		DDPA + DSRBACA	RBAC
1	100	6	5
2	200	10	9
3	300	15	13
4	400	19	17
5	500	20	22

values analysed to test clients accessing the cloud storage and access denial in the proposed framework. For this reason, we have attempted an alternate number of client demands, for example, 100, 200, 300, 400 and 500 of every five analyses. Table 1 shows the client demands denied examination. Additionally, the solicitations incorporated the certifiable and the malignant client demand with an extent 19:1.

In Table 1, the results show that the Dynamic Spatial RBAC Algorithm and the ordinary RBAC model were compared in confining the cloud user quantity, furthermore gives over 90% location and anticipation precision. This is on the grounds that the utilization of worldly and spatial requirements while settling on a choice. Successful and made sure about proposed deduplication is utilized.

The performance of clients' request wherein the quantity of approved clients in the DSRBAC algorithm is shown in Fig. 4.

In Fig. 4, the DSRBACA authorization access is less than the existing RBAC algorithm. In addition, cloud users less than 5% in the current framework were denied access, and subsequently the RSA encryption algorithm upgrades the security. This

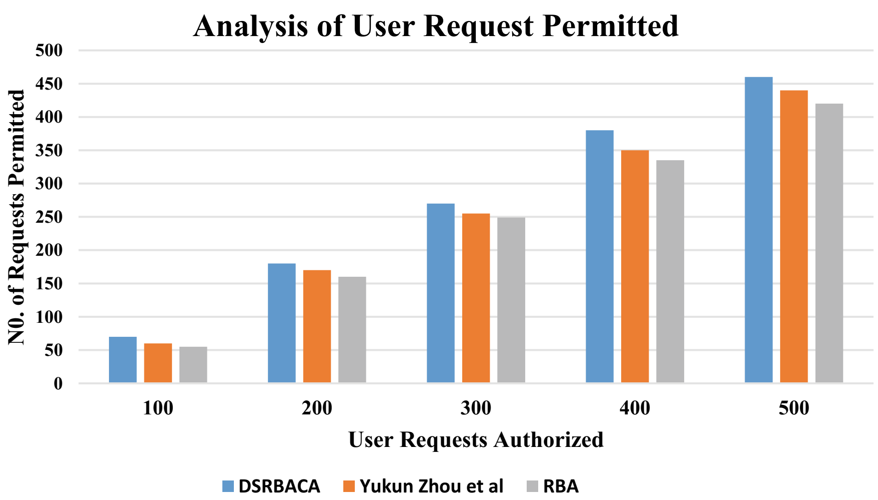


Fig. 4 Analysis of permission on user request by RBAC and DSRBACA

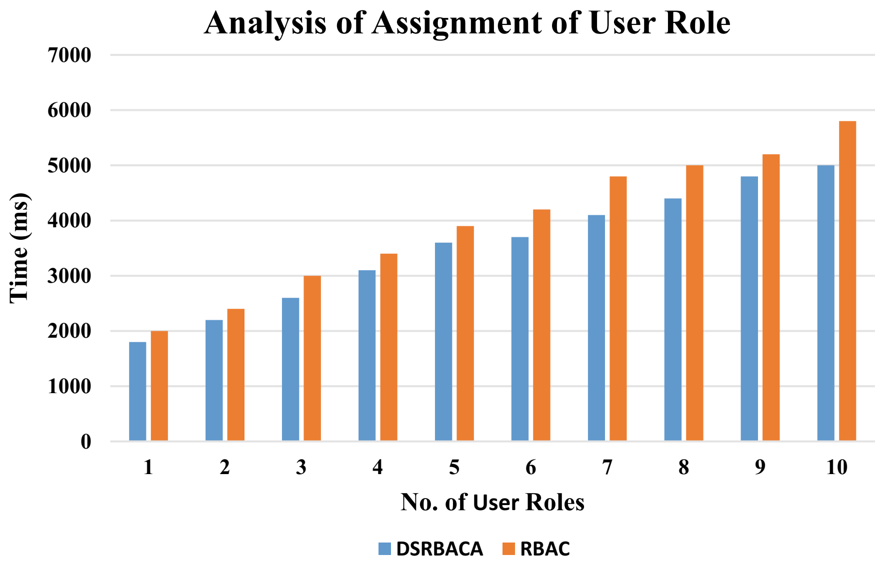


Fig. 5 Analysis of assignment of user role

is a direct result of the utilization of existence requirements viably for distinguishing the irregular clients.

The analyses directed needs 959 ms to make 10 roles, 9 edges are incorporated in 844 and 711 ms to concede 10 permissions for every one of the t roles, and 3,384 ms to dole out 50 clients to every job. The time involved for every one of the basic activities is in the range 68–120 ms. The time taken to dole out ten clients with ten roles is shown in Fig. 5.

DSRBACA and RBAC were compared in terms of the time taken to process the constraints in Fig. 5, graph shows that DSRBACA takes less time than RBAC.

Figure 6 examines the approval time examination of the proposed DSRBACA, RBAC and renouncement algorithm [18]. Here, we have led five examinations with an alternate number of documents, for example, 100, 200, 300, 400 and 500 MB.

Figure 6 depicts the enhanced security performance of the DSRBACA, standard RBAC and Zhou et al. [19].

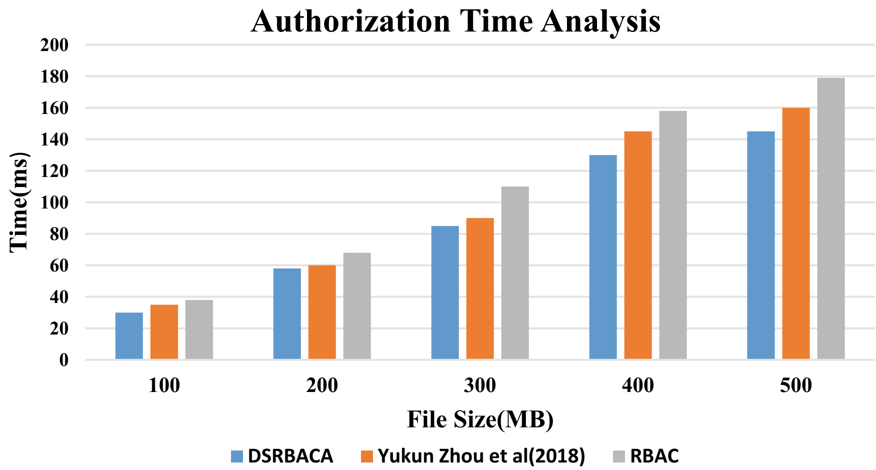


Fig. 6 Authorization time analysis

6 Summary

Despite the fact that all the essential activities are insufficient for cloud storage and recovery. A large quantity of data in cloud storage application is distributed in numerous locations. These drawbacks have been overwhelmed by the recently proposed Dynamic Spatial RBAC Algorithm is dynamic dependent on time and location. This mix of location and time are useful for additional improvement of the cloud data security. DSRBACA has constrained the access of client's information or records that in the hybrid cloud storage system. In correlation with the RBAC model in limiting the quantity of cloud clients, the proposed DSRBACA performs well and furthermore gives over 90% location and avoidance precision.

Moreover, approved cloud users less than 5% were denied access in the current framework and consequently the RSA encryption algorithm upgrades the security with better confidentiality of data. Generally speaking, the security dangers in the hybrid cloud systems were improved utilizing cryptography algorithms.

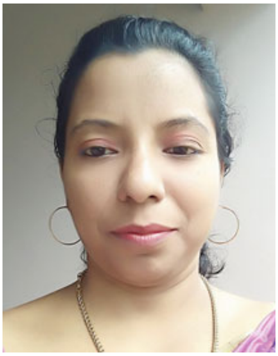
References

1. Praveena, D., Rangarajan, P.: Analysis of trend, service and deployment models in cloud computing with focus on hybrid cloud and its implementation. Res. J. Appl. Sci. **9**(4), 181–186 (2014)
2. Bellare, M., Keelveedhi, S., Ristenpart, T.: Message locked encryption and secure deduplication. In: Proceedings in 32nd Annual International Conference on Theory Application Cryptographic Technology, pp. 296–312 (2013)
3. Yuan, H., Chen, X., Jiang, T., Zhang, X., Yan, Z., Xiang, Y.: DedupDUM: secure and scalable data deduplication with dynamic user management. Inf. Sci. **456**, 159–173 (2018)

4. Praveena, D., Rangarajan, P.: Improving the security in web application cloud service Karatsuba Montgomery multiplier ECC algorithm. J. Web Eng. **17**(6), 3078–3094 (2018)
5. Thanga Ramya, S., Arunagiri, B., Rangarajan, P.: Novel effective X-path particle swarm optimization based deprived video data retrieval for smart city. Clust. Comput. 13085–13094 (2019)
6. Shete, A., Patil, B.M.: Deduplication in hybrid cloud with secure data. Int. J. Comput. Appl. **148**(8), 32–36 (2016)
7. Raji, L., Thanga Ramya, S., Thilagavathy, A.: Reducing cloud storage space consumption using PRCR. Int. J. Innov. Technol. Explor. Eng. (TM) **46**, 287–289 (2019)
8. Zhou, L., Varadharajan, V., Hitchens, M.: Achieving secure role-based access control on encrypted data in cloud storage. IEEE Trans. Inf. Forensics Secur. **8**(12), 1947–1960 (2013)
9. Akhila, K., Ganesh, A., Sunitha, C.: A study on deduplication techniques over encrypted data. Procedia Comput. Sci. **87**, 38–43 (2016)
10. Meyer, D.T, Bolosky, W.J.: A study of practical deduplication. ACM Transactions on Storage (TOS) **7**(4), 1–20 (2012)
11. Gapat, P., Khillare, S., Khiste, A., Pise, R., Gapat, P.: Securing data deduplication on hybrid cloud using asymmetric key algorithm. J. Innov. Res. Sci. Technol. **2**, 759–763 (2017)
12. Praveena, D., Rangarajan, P.: A machine learning application for reducing the security risks in hybrid cloud networks. Multimedia Tools Appl. **79**(7), 5161–5173 (2020)
13. Qiu, M., Gai, K., Singham, B.T., Tao, L., Zhao, H.: Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry. Future Gener. Comput. Syst. **80**, 421–429 (2018)
14. Prema Rajeswari, L., Arputharaj, K.: An active rule approach for network intrusion detection with enhanced C4.5 Algorithm. Int. J. Commun. Netw. Syst. Sci. **1**(4), 314–321 (2008)
15. Yan, Z., Wang, M., Li, Y., Vasilakos, A.V.: Encrypted data management with deduplication in cloud computing. J. Cloud Comput. **3**(2), 28–35 (2016)
16. Zhenhua, L., Yaqian, K., Chen, L., Yaqing, F.: Hybrid cloud approach for block-level deduplication and searchable encryption in a large universe. J. China Univ. Posts Telecommun. **24**(5), 23–34 (2017)
17. Helmi, A.M., Farhan, M.S., Nasr, M.M.: A framework for integrating geospatial information systems and hybrid cloud computing. Comput. Electr. Eng. **67**, 145–158 (2018)
18. Li, J., Li, Y.K., Chen, X., Lee, P.P.C., Lou, W.: A hybrid cloud approach for secure authorized deduplication. IEEE Trans. Parallel Distrib. Syst. **26**(5), 1206–1216 (2014)
19. Zhou, Y., Dan Feng, Yu., Hua, W.X., Min, F., Huang, F., Zhang, Y.: A similarity-aware encrypted deduplication scheme with flexible access control in the cloud. Future Gener. Comput. Syst. **84**, 177–189 (2018)

Dr. D. Praveena B.E., M.E., Ph.D., is an Assistant Professor in Department of Information Technology, RMD Engineering College, since August 2006. She has completed his Ph.D. (Hybrid Cloud Security) at College of Engineering, Anna University in the year 2019. She has been in the teaching profession for the past 15 years and has handled UG programmes and published 7 papers in various International Journals and Conferences.

Dr. S. Thanga Ramya B.E., M.S. (by Res), Ph.D., is an Associate Professor in the Department of Information Technology, since June 2008. She obtained her B.E. (CSE) from Dr. Sivanthi Aditanar College of Engineering and M.S. by Research (ICE) from Anna University, Chennai. She has obtained her Ph.D. in Information and Communication Engineering from Anna University, Chennai, in 2017. She has been in the teaching profession for the past 18.5 years and has handled both UG and PG programs. Her areas of interest include programming languages, database management and data mining. She has published 13 papers in various International Journals and Conferences. She has also published four text books for the benefit of B.E./B.Tech. students of Anna University. She has got Oracle java international certification, IBM RAD, TIVOLI and DB2 certification and obtained her PRP certification from Wipro. She secured bronze medal from Infosys for the campus connect program. She is the life member of ISTE.



Dr. V. P. Gladis Pushparathi B.Tech., M.Tech., Ph.D., is an Associate Professor in Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, since Dec 2016. She obtained her B.Tech. from Cape Institute of Technology, Tirunelveli and M.Tech. from M.S. University Tirunelveli. She has obtained her Ph.D. in Information and Communication Engineering from Anna University, Chennai, in 2016. She has been in the teaching profession for the past 14 years and has handled both UG and PG programs. She has completed major and minor funded projects. She has published 21 papers in various International Journals and 78 papers in international Conferences. She has published more than 15 Indian patents. She has also published three text books for the benefit of B.E./B.Tech. students of Anna University. She has got best faculty award from Neghru Group of Institutions. She is the life member of ISTE, IACSIT, IAENG and ISRD. She is the reviewer for Journal of computer and Electrical Engineering/Elsevier and Journal of 3D Research. Her field of Interest includes Digital Image Processing, Soft Computing, and Data mining.

Mr. Pratap Bethi working as General Manager (SAP & IT) in OPG Power Generation Pvt. Ltd. He has around 17 years of experience in SAP and other verticals. He completed Bachelor's degree in Technology (B.Tech.) with the specialization in Mechanical Engineering from J.N.T.U. in April' 2003. His area of interest is ERP/Integration of third party systems with SAP and Cloud Technologies.

Mr. S. Poopandian is a principal consultant in Infosys Pvt., Ltd., He has 16 years of experience in SAP related project (Building Global Template, Implementation of Product costing, COPA International Rollout Programs, AMS Lead), SAP Implementations, Rollouts and Upgrade. He has working experience in multicultural environments (Asia, US, Europe, Africa, Asia). He is SAP Management Accounting Certified Solution Consultant ECC 6.0. He Worked on S4 Hana on POC for Various clients. He is having expert knowledge in SAP RAR solution for IFRS 15 Compliance. He has done couple of India CIN Implementation and 3 India GST Implementation for Global Automobile Majors, Wind power energy producers and Car Chassis Manufacturers. His area of interest mainly focuses on cloud computing.

Analysis of Long Short Term Memory (LSTM) Networks in the Stateful and Stateless Mode for COVID-19 Impact Prediction

Vinayak Ashok Bharadi and Sujata S. Alegavi

Abstract Machine learning (ML) has become a trending domain over the past few years, the accessibility of Graphical Processing Units (GPUs), Tensor Processing Units (TPUs) have given impetus for the same. Various applications like speech and face recognition, natural language processing, text analytics, big data analytics, regression analysis, pattern recognition and classification are based on the machine learning concept. Regression analysis evaluates the impact of a set of variables among themselves as well as the final formulation. Using this fitting of a particular theory for the real-world inputs can be evaluated. In this chapter regression analysis is performed on the COVID-19 data to predict the next values of the parameters. The Long Short Term Memory Networks (LSTMs) are used here for the prediction task, the LSTMs come under a special category of Neural Networks known as Recurrent Neural Networks (RNNs) which are used for this prediction task. The stateless and stateful implementation of LSTMs are designed and their performance is evaluated. The details of stateful and stateless architecture and their implementation in Keras framework is presented here. The results indicate that the LSTMs have better performance as compared to the RNNs.

1 Introduction

Machine learning is a set of methods for creating models automatically from the input data. Machine learning algorithms are at the core of it, they do the learning part for the automated model generation. There are various types of algorithms like classification, regression, supervised, unsupervised etc. The efficiency and accuracy

V. A. Bharadi (✉)
Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India
e-mail: vinayak.bharadi@famt.ac.in

S. S. Alegavi
Department of Electronics Engineering, Thakur College of Engineering and Technology, Mumbai, India
e-mail: sujata.dubal@thakureducation.org

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_8

of the machine learning models depends on the problem that an individual is working on. Artificial Intelligence (AI), Machine learning (ML), and Deep learning (DL) have proven useful and energizing for the organizations and businesses which are metrics-driven.

The bottleneck in the data value chain is the capacity to consume the data perform the analytics and use it in the intelligent way for the further development of business intelligence. This makes a great deal for business and process professionals to work with AI, machine learning, and deep learning in congruence with the business frameworks and concepts. This will drive the data scientists to define problems and hypotheses which are part of the above-mentioned frameworks and concepts, and further to use AI, ML and DL to find patterns or the data insights and to test hypotheses which are otherwise computationally expensive or difficult to program [1].

Conventional Programming is focussed on instructing a computer the things to do in an upfront manner. A straight line or a different linear function that is linear in its parameters such as a polynomial, to numeric data is fit to the input data when the Linear regression is used. Typically, this is achieved by performing matrix inversions to minimize the squared error between the fitted line and the input data. For the problems that can't be solved with the deterministic methods, Nonlinear regression is used; This fits the curves that are not linear in their parameters to data, and its bit complicated also as it uses some kind of iterative minimization process and the variation on the method of steepest descent also.

Machine Learning provides more complex solution as compared to the once mentioned above, mathematical modelling for the solutions may also be quite difficult. In the literature one can find two main types of problems that are commonly addressed by machine learning; they are regression and classification. The Regression analysis is done for the numeric data and for non-numeric data classification is performed. In this chapter the regression analysis on the COVID-19 pandemic data is performed which uses a special type of neural networks called as Long Short Term Memory (LSTM) Networks. The LSTMs are a special type of networks in the category of Recurrent Neural Networks [2].

2 Recurrent Neural Networks

Recurrent Neural Network (RNN) operates on sequences of vector which are not fixed unlike the Feed Forward Neural Networks (FFNN) or the Convolutional Neural Networks (CNNs) or, which allows various possibilities to operate at the input side, output side or in the network where the number of layers are not fixed. The FFNN are represented from left to right, beginning from the input layer to the output layer and in between having a certain number of hidden layers. The weights are updated from the prediction errors using back propagation algorithm. This is called the learning process and it is an iterative mechanism, for every iteration the network starts from scratch, considers current inputs only and does not consider previous inputs or state of

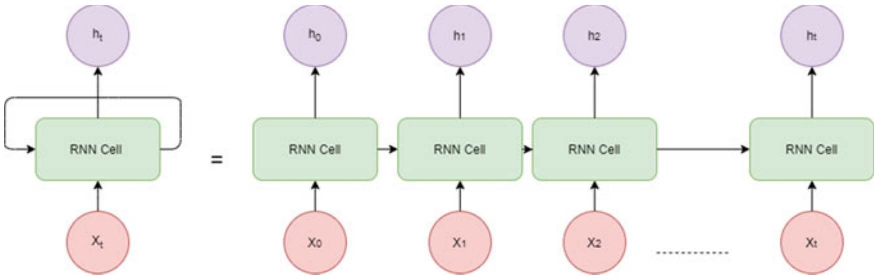


Fig. 1 Conceptual representation of a typical RNN cell

the network. This is very much suitable for applications like classification, clustering etc. where the temporal information of the input is redundant. However, consider a problem of translating a given input text or prediction of a value in a sequence of number, this brings notion of time in picture and the FFNN can't process this data effectively [3].

To address this problem of sequence processing; Recurrent Networks are designed. The word "Recurrent" means something happening repeatedly for certain amount of time. In this type of networks each time the model receives a new input from the sequence data it will process it and this same step will be repeated for the rest of the inputs consequently and the processing will be recurrent. Further as it was stated above that the RNNs have notion of time; it will remember the previous inputs and state of the network and store it in memory so when we process the subsequent inputs, the network has the context of the past to process the current data.

As discussed earlier a typical RNN can be seen in Fig. 1. The recurring inputs to the hidden layer which are also the output from a hidden layer is fed back to the cell itself. The sequence of images on right show the same layer unfolded into the time steps as the output of one-time step are fed as input to the next step. Because of this RNNs have deep in time nature inherently, as their current hidden state is a function of all the past hidden states.

RNNs have an added advantage from depth in space; that is from adding multiple recurrent hidden layers on top of each other, like the feedforward layers are arranged in the conventional deep networks [4]. With the addition of multiple layers, you enable the RNNs to analyse more complex structures.

2.1 From RNN to LSTM

Due to the recurrent structure of the network, the memory that is being produced during the training phase is limited. The network weights go on becoming infinitesimally small or very large due to the feedback mechanism. This make the network unable to retain the temporal context of the data and it is called the vanishing gradient problem [5]. All the algorithms used in the training phase have fallen victim to either vanishing or exploding gradients, which results in the failure of the model. The model fails to learn long term sequential dependencies in the data. To overcome the problem of vanishing and exploding gradients, Long-Short term memory popularly known as LSTM is been designed [6]. When some information is no longer relevant, the LSTM learns the rate at which the value decays which is stored in the memory cell; this is done through a special construct called as the forget gate. LSTM's are far more complex than regular RNN models. LSTM's are basically designed to understand and remember long term dependencies.

LSTM's being a popular tool in RNN used for prediction which uses past and current data to forecast the future data. There are many fields in which LSTM' s is used successfully like image classification, image captioning, sentiment analysis, machine translation and video classification.

3 LSTM Architecture

LSTM has three significant gates also referred as cell states, as well as one global cell state, this is shown in Fig. 2 Many researchers have tried to change the basic architecture of the LSTM's over the time to suite for various applications. A single LSTM cell can be seen below, in practice such multiple cells are stacked together to generate LSTM network [7].

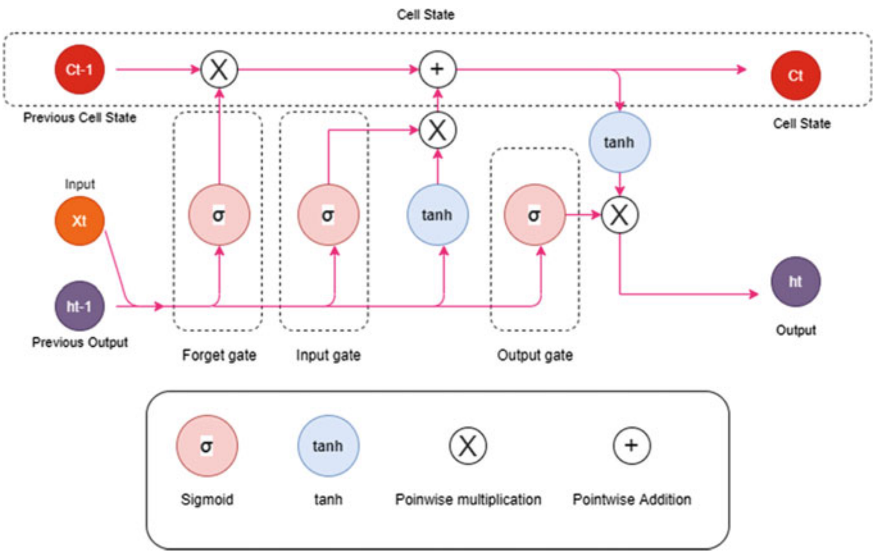


Fig. 2 Conceptual diagram for long short term memory network cell

3.1 Various Gates in the LSTM Architecture

The architecture of the LSTM consists of three cell states known as gates. The input gate, the forget gate and the output gate. They are described in Table 1.

3.1.1 Input Gate

Input gate is to update the current cell state. The current input and previous hidden state are passed into the sigmoid function. The sigmoid function is used to resolve which value will be updated to 0 or 1. Here, '0' indicates not important whereas '1' indicates important. The hidden state and current inputs are also passed into the tanh function to make the adjust the values between '1' and '−1' to regulate the network. The tanh output is multiplied by the sigmoid output and the sigmoid output will now decide which value from the tanh output is important to keep.

3.1.2 Forget Gate

The forget gate plays a major role in deciding whether to keep the information or to throw it away. If the value of forget gate ft becomes '1' then information is kept in the cell state and if the value of the forget gate becomes '0', then the information is thrown away from the cell state. The bias bf that is attached with the forget gate can be increased which in turns increases the performance of the LSTM network.

Table 1 Gates in a typical LSTM and their functions

Type of entity	Role	Equation
Input gate	Responsible for adding information to the cells	$\mathrm{i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i)}$
Forget gate	Removes information that is no longer necessary for the completion of the task. This step is essential to optimize the performance of the network	$\mathrm{f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f)}$
Output gate	The next hidden state is determined by it. Selects and outputs necessary information	$\mathrm{o_t = \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o)}$
Cell state	State of the LSTM cell	$\mathrm{i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i)}$ $c_t' = \tanh(W_{ch}h_{t-1} + W_{cx}x_t + b_c)$ $\mathrm{c_t = f_t \odot c_{t-1} + i_t \odot c_t'}$
Output	Output of LSTM	$h_t = o_t \odot \tanh(c_t)$

3.1.3 Cell State

To calculate the cell state, the current cell state is pointwise multiplied by the forget vector. This remover some of the values in the cell state, as these values get multiplied by '0'. The new cell state is generated by, the pointwise addition of previous cell state and output of the input gate and pointwise multiplication of tanh function output. This part adds the notion of memory to the LSTM cell.

3.1.4 Output Gate

The next hidden state is decided by the output gate. The information of the previous inputs is stored in the hidden state which is used for predictions. The current input and the previous hidden state are passed into the sigmoid function. The tanh function is fed with the modified cell state. To decide what information the hidden state should carry, tanh output is multiplied with the sigmoid output. The hidden state is formed by the output. Further we carry the hidden and cell states, which are newly formed to the next time step.

There are several versions of these LSTMs which have been developed over the time. The latest in RNN is called the Gated Recurrent Unit (GRU). The GRU has been designed using an update gate in its structure, which is a grouping of an input gate and a forget gate [8]. The GRU model was designed to simplify the architecture of the LSTM. In this chapter the LSTM and its stateful and stateless implementations will be discussed.

3.2 Stateful and Stateless LSTM

There are two modes in a LSTM network Stateful and Stateless mode. In the stateless mode the LSTM Updates parameters for each batch. The LSTM is organized as layers of Cells or neurons. Same way it is referred in all other neural networks. So even if CNN, convolutional neural networks, or other networks are considered, there is always an input layer. Then hidden, output layers are arranged. And this state of neurons in the hidden layers, called hidden state. So, it's now known that in a stateless LSTM, everything starts from beginning from batch to batch. Hence one batch was past, then everything is going to be reset and initialized with zeros and then it starts again and again. In sentence prediction, suppose there is a sentence in one batch and other sentence in other batch and the data between these two batches is completely independent of each other in such a case the stateless LSTM networks are used.

In a stateful neural network, the last output of the hidden state and of the cell states, cell state again, this is the cell memory, from batch one is the input for the batch two. So, it memorizes what it has learned in the batch one, it takes it over to the batch two [9]. Whenever there is discussion about sequence to sequence prediction and huge data (observations) is under the context, which is split into different batches. The data in one batch is dependent on data in the other batch, so there is a dependency inside of this data and between different timesteps which represents time series and hence use of stateful LSTM's becomes more important in such a case. So, depending on the different applications stateful or stateless LSTM networks are selected.

In sequence to sequence prediction as there is a dependency of data between different time steps hence, use of stateful LSTMS's becomes much preferable.

4 LSTM Research

In one of the sequences to sequence based prediction the authors focused on early detection and prediction of Influenza using LSTM's [10]. This work suggests use of Seq2Seq with attention model using google trends data which allows to predict over multiple time steps covering many weeks. In another implementation the researchers suggested use of multi-channel LSTM neural network that can draw multiple information from different types of inputs, by adding attention layer forecasting is further improved [11]. Trang Pham et al. have proposed Deep care which stores the medical records, previous illness history, current illness states and predicts the future medical outcomes based on LSTM's. Sequence to Sequence architecture is also used in predicting MiRNA sequence from MRNA sequence using stacked LSTM's with proper dropout regularization [12]. Rajkumar Chakraborty and Yasha Hasija in this paper suggested that by predicting miRNA from MRNA many mutations which are caused in human genome can be decoded which results in various disease in humans [13]. Even chronic neurodegenerative disease like Alzheimer's Disease (AD) can be predicted using LSTM efficiently as suggested by Hong et al. [14]. As this disease develops in the temporal region of the brain, hence this data for the patient are potentially meaningful for predicting the development of the diseases. In this paper instead of time series data, time step data is used by data pre-processing pipeline and based on this data LSTM time sequence model is built to predict the AD at an early stage. Jiyun Zhou and et al. suggested Prediction of DNA-Binding Residue from Protein Sequence using LSTM along with Ensemble Learning [15] which uses two layered model. The first layer consists of LSTM layer that uses a bi-gram model the pairwise relationships between residues and then learns feature vectors for all residues and the second layer Ensemble Learning based classifier introduced to tackle the data imbalance problem in binding residue predictions.

In the recent crisis of Covid-19 which is declared as a pandemic by WHO due to which the entire world has gone into lockdown phase, methods of identification of humans suffering from this disease has become quintessential and of utmost urgency [16]. Sequence to Sequence prediction with LSTM model can be used to predict human beings which may suffer from this virus in the future. Taking into account the current and previous database the model can learn the trends in the data and predict possible outbreak regions as well.

Sequence to Sequence prediction with LSTM model can be used to predict human beings which may suffer from this virus in the future. Taking into account the current and previous database the model can learn the trends in the data and predict possible outbreak regions as well.

5 COVID19-Prediction Problem

In early 2020, a new virus began to appear in news all over the world because of the unparalleled speed of its transmission and infection. This virus is believed to be originated in a food market in Wuhan, China, in December 2019. Since then it has reached countries as distant as the USA and the Britain and gulf countries [16]. The virus is officially named as SARS-CoV-2 has been the cause of 10.2 Million of infections globally, causing 502 thousand deaths. The United States is the country most affected with 2.6 Million infections and 128 thousand death as on July 2020.

The SARS-CoV-2 infection causes the disease called as is called COVID-19, which stands for coronavirus disease 2019. Figure 3 Shows the spreading of COVID-19 from china to other parts of the world. On the X axis it shows the timeline and on the Y axis it shows the cumulative infections of the virus.

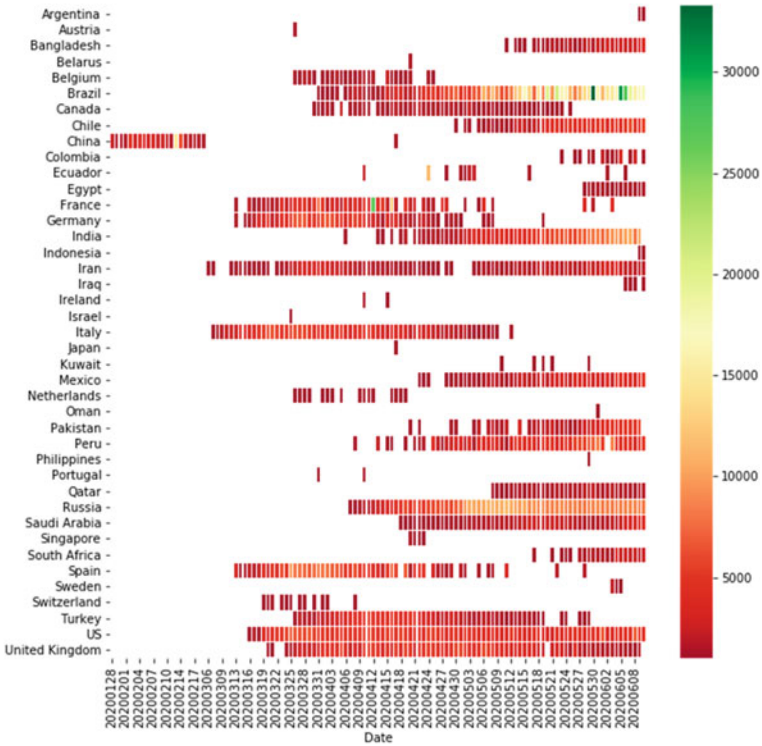


Fig. 3 Heatmap indicating the spread of COVID-19 pandemic from China to other countries

In this work the COVID19 daily status used as the input, this data is made available by John Hopkins University. The data till 27th May is considered as input and it is given to the following type of neural networks

1. Recurrent Neural Network
2. Stateful LSTMs
3. Stateless LSTMs

The prediction is compared with the actual figures.

The input data has following counts available for 115 days for COVID19 Pandemic

1. Global Confirmed Cases
2. Global Deaths
3. Global recovery
4. Confirmed Cases in India
5. Deaths in India
6. Recovery in India
7. Confirmed Cases in USA
8. Deaths in USA
9. Recovery in USA

The Plot for the data points can be seen in Fig. 4.

The nine sequences mentioned above are correlated and they are taken as input in univariate as well as multivariate mode. They are fed to RNN as well as LSTMs and next figures are predicted. These predictions are compared with actual figures

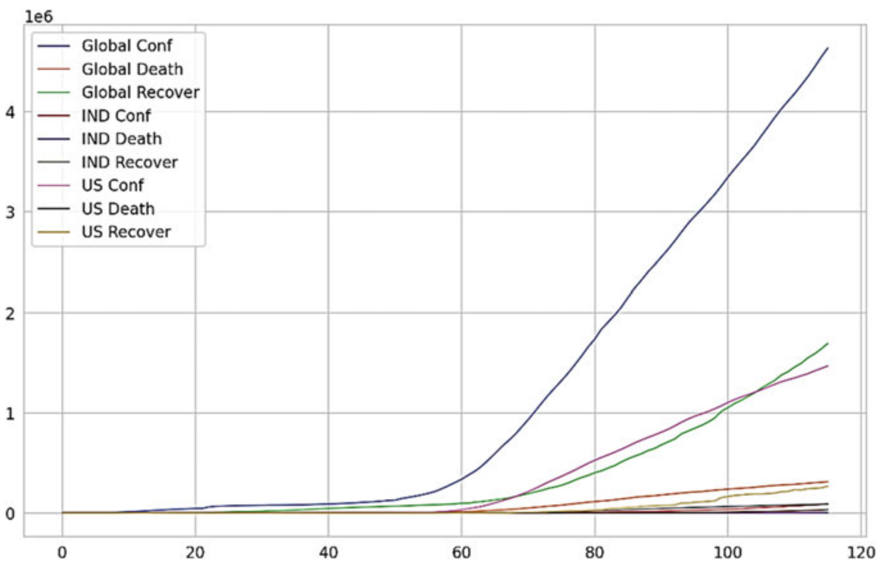


Fig. 4 COVID19 data plot (actual figures scaled by 100,000 on Y-axis)

for evaluation of error. For the implementation of LSTMs, the Keras framework [17] is used.

Francois Chollet a Developer/Researcher at Google AI created and developed the Keras framework. Initially, Theano was the default backend of Keras and until v1.1.0. Keras started supporting TensorFlow as a backend. Later on, the TensorFlow became the most popular and the default backend starting from the release of Keras v1.1.0. With the announcement of the TensorFlow 2.0 in June 2019, Google declared Keras as the official high-level Application Programming Interface (API) of TensorFlow for quick and easy model design and training. The Keras v2.3.0 was officially released on September 17th, 2019, this is the first release of Keras which is in sync with `tf.keras` [18].

6 Defining Models in Keras

Keras is a Neural Network library coded in Python that runs on top of Theano or TensorFlow. Keras is an open source software. Keras uses a library or "Backend" to run the low-level computation. This makes Keras a high-level API wrapper for the low-level API, using this approach Keras is capable of running on top of frameworks such as The Microsoft Cognitive Toolkit (CNTK), Theano or the default TensorFlow.

Keras High-Level API handles the way the Neural Network models are deployed, define layers, or setting up multiple input-output models. Keras also compiles the model with loss and optimizer functions at this level, further it trains and fits the model. The processes related to Low-Level API such as making the computational graph, making tensors or other variables are handled by the "backend" engine.

The actual neural network model is represented by Keras model. The models can be created in two modes using Keras; First one is simple and easy to use Sequential API and next is more flexible and advanced Functional API. The sequential API of Keras enables one to create the model layer-by-layer for the given problems at most of the times. If the models share layers or they have multiple inputs or outputs it is not possible to define model in sequential mode in Keras. In this type of models, the layers are arranged in a sequential manner, hence the name 'Sequential". Majority of the ANN also have their layers arranged in sequential manner and the data flow from one layer to other layer in a linear manner till it reaches to the output layer.

On the other hand, the functional API is much more flexible; it allows to define models having layers connecting to more than just the previous and next layers. Additionally, the layers can be connected to any other layer. It empowers creation of complex networks such as residual networks and Siamese networks. For current discussion the sequential mode is used.

6.1 Five Step Lifecycle

To work with an ANN model in Keras one has to go through the five-step life cycle. The steps are listed below

1. Define the Model.
2. Compile the Model.
3. Fit the Model.
4. Evaluate the Model.
5. Make Predictions with the Model.

A stateless model in Keras will be implemented through the above mentioned five steps.

6.1.1 Define the Model

The LSTM recurrent layer consists of memory units is called as LSTM (). This is a fully connected layer which follows LSTM layers and is used for giving an output prediction layer called as Dense (). The first hidden layer in the LSTM network should define the number of inputs to expect i.e. the shape of the input layer. Input must be three-dimensional (3D), consisting of the samples, time steps, and features as detailed below.

- **Samples**. These are the reading or the rows in your data.
- **Time steps**. These are the set of previous observations for a feature. For a sample it indicates the time scale past of that feature.
- **Features**. These are columns in your data or the attributes of the data.

So, the input has to be reshaped in collection of [Sample, timesteps, Features] to be given as input to the LSTM.

As discussed in section five, the data is available as a 2D matrix, comprising 115 rows and 9 columns. For the current prediction the prediction problem is modelled as a multivariate and multistep sequence prediction. It will take multiple variables as an input and will predict multiple variables as an output, each input sample consists of multiple time steps.

As the input data is having very high numeric values, it is better to scale the same, hence the min-max scalar transformation is used to scale it to 0–1 range. Then to add the time step to each sample, the data is reshaped. This is like for each row in the matrix, the past N readings are considered and grouped together to form a composite sample entity consisting of that many rows and it makes one such sample. Out of 115 samples 88 samples are considered for training and 19 for testing, further we have N = 4 (number of time steps) hence the initial shape of input is [88, 9] after reshaping with 4-time steps it becomes [88, 4, 9].

LSTM networks are defined in Keras as a sequence of multiple layers. The container for these layers is the Sequential class. The LSTM recurrent layer comprised of memory units is called LSTM (). A fully connected layer that often follows LSTM layers and is used for outputting a prediction is called Dense ().

```
# define model
model = tf.keras.Sequential()
model.add(layers.LSTM(n_samples, activation='relu',
          return_sequences=True, input_shape=(n_steps, n_features)))
model.add(layers.LSTM(n_samples, activation='relu',
          return_sequences=True, input_shape=(n_steps, n_features)))
model.add(layers.LSTM(n_samples, activation='relu'))
model.add(layers.Dense(output))
```

6.1.2 Compile the Model

The compilation is a precompute step for the defined neural network model. It is always necessary after defining a model. The algorithm for the training, its parameters, metric to calculate the loss in the training process as well as accuracy metrics are defined here and the model is compiled. A typical example is as follows:

```
algorithm = tf.keras.optimizers.RMSprop()
model.compile(optimizer=algorithm, loss= tf.keras.losses.mean_abso-
lute_error, metrics=['accuracy'])
```

The compilation of the network gives the following output:

```
Using TensorFlow backend.
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 1, 91)             36764
_________________________________________________________________
lstm_2 (LSTM)                (None, 1, 91)             66612
_________________________________________________________________
lstm_3 (LSTM)                (None, 91)                66612
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 184
=================================================================
Total params: 170,172
Trainable params: 170,172
Non-trainable params: 0
_________________________________________________________________
None
```

It can be seen that the LSTM layer has 170,172 parameters. This is calculated based on the number of inputs = 9 (no of features) and the number of outputs (91 for the 91 units in the hidden layer), as follows:

```
n1 = 4 * ((inputs + 1) * outputs + outputs^2)
n1 = 4 * ((9 + 1) *91 + 91^2)
n1 = 4 * 9191
n1 = 36764
```

The second and third layer is having 91 inputs from the previous layers, hence

```
n2 = 4 * ((inputs + 1) * outputs + outputs^2)
n2 = 4 * ((91 + 1) *91 + 91^2)
n2 = 4 * 16653
n2 = 66612


n3 = 4 * ((inputs + 1) * outputs + outputs^2)
n3 = 4 * ((91 + 1) *91 + 91^2)
n3 = 4 * 16653
n3 = 66612
```

Further the fully connected layer only has 91 parameters for the number of inputs (91 for the 91 inputs from the previous layer), number of outputs (2 for the 2 neurons in the layer), and the bias.

```
n4 = inputs * outputs + outputs
n4 = 91 * 2 + 2
n4 = 184
```

`total parameters = n = n1+ n2+ n3+ n4  = 170,172`

6.1.3 Fit the Model

After compilation the network can be fit. In this step the network adapts the weights on a training dataset. This requires a training dataset which specifies the input and output matrices. The backpropagation algorithm is used for the training. The network is optimized as per the parameters such as the optimization algorithm and loss function specified in the compilation step.

The backpropagation is performed iteratively for a specified number of epochs or exposures to the training dataset. The data is passed in batches of the input and output in each epoch.

The batch size indicates the number of patterns that the network is exposed to before its weights are updated in an epoch. The batch size also brings efficiency optimization, ensuring only enough number of input patterns are loaded into memory at a time. The fit process returns a history object that includes a summary of the performance of the model during the training. If then validation is also performed after fitting then the history object includes training and validation the loss and any additional metrics specified when compiling the model. This is done as follows:

```
history = model.fit(X, Y, validation_split=0.20, batch_size=5,
epochs=200, verbose=0)
```

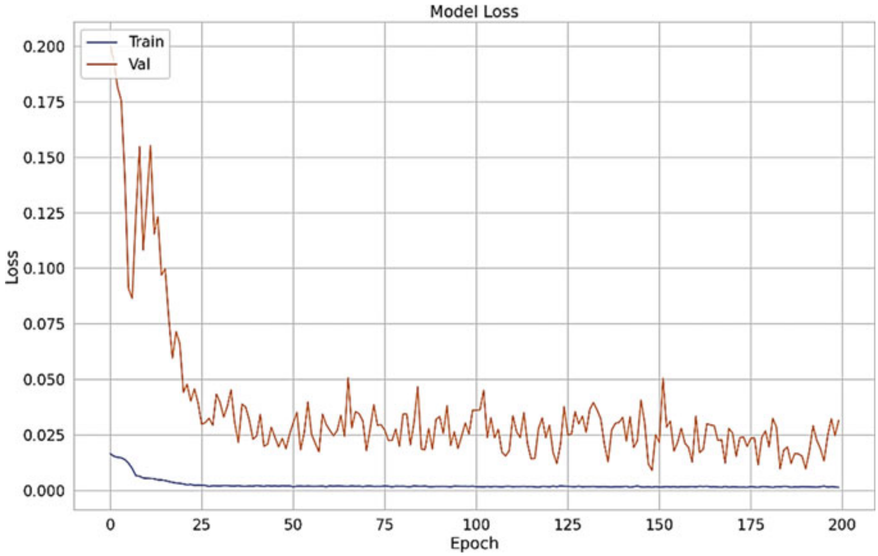


Fig. 5 Output of model fitting—training and validation plots for loss

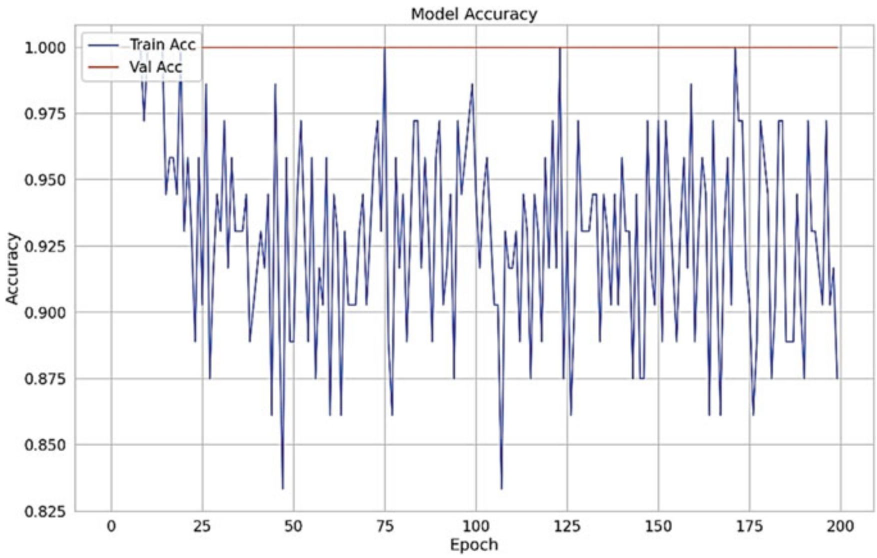


Fig. 6 Output of model fitting—training and validation plots for accuracy

The training and Validation loss as well as accuracy is show in Figs. 5 and 6, this is a sample output and performance can be further improved by modifying the network definition and parameters.

6.1.4 Evaluate the Model

The network performance is evaluated on a separate dataset, this is different as compared to the training data. This gives an estimate of the performance of the network towards making predictions for unseen data in the future. The command for evaluate and the output is given below

```
history2 = model.evaluate(Xt,Yt,batch_size=5)
```

```
5/5 [==============================] - 0s 2ms/step - loss:
0.2095 - accuracy: 1.0000
```

6.1.5 Make Prediction

Once the network is satisfactorily fit and evaluated, it can be used to make predictions on new data. This is done by calling predict () function on the model with an array of new input patterns. The output for then prediction is given below in Fig. 7.

7 LSTM State Management

Every LSTM memory unit maintains its accumulated internal state. This internal state has to be carefully managed for specific sequence prediction problem not only during the training of the network but also when making predictions.

The internal state of all LSTM memory units which are the part of network is by default reset after each batch. This is the instance when the network weights are updated. It indicates that configuration of the batch impacts

- The number samples which are processed before an update, this is the scope of the learning in view of a batch.
- Learning speed, or the frequency of weight update.
- The frequency of resetting internal state.

In Keras it is possible to flexibly decouple the resetting of internal state from updates to network weights. This is done by defining a stateful LSTM layer by setting the stateful argument on the LSTM layer to True.

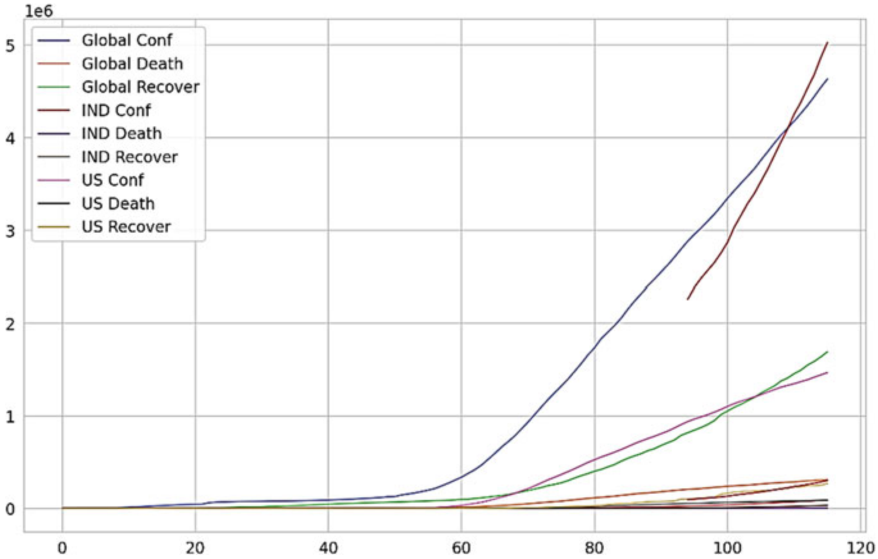


Fig. 7 The COVID19 cases prediction by regular stateless LSTMs—shown in red markings over the actual values

In the implementation of stateful LSTM layers the batch size must be defined as part of the input shape. This has to be done in the definition of the network by setting the batch input shape argument. Further, the batch size must be a factor of the number of samples in the dataset used for the training of the network.

The batch input shape argument t is a 3-dimensional tuple formed as batch size, time steps, and features. Consider the following example, a stateful LSTM can be defined to be trained on a training dataset with 200 samples, a batch size of 20, and 10 time steps for 2 features, as follows.

```
# 1. define network
batchSize = 20
timeSteps = 10
features = 2
model = tf.keras.Sequential()
model.add(layers.LSTM(2, stateful=True, batch_input_shape=(batchSize,
 timeSteps, features)))
model.add(layers.Dense(1))
```

A stateful LSTM does not resets the internal state at the end of each batch. On the other hand, a fine grained control over when to reset the internal state by calling the reset_states() function is available. If it is required to reset the internal state at the end of each single epoch, it can be done as follows:

```
for i in range(100):
model.fit(X, Y, epochs=1, batch_input_shape=(20, 10, 2))
model.reset_states()
```

Care should be taken to use same batch size as used in the definition of the stateful LSTM while making predictions.

```
predictions = model.predict(X, batch_size=20)
```

The internal state in LSTM layers are accumulated during the network evaluation as well as while making predictions. Therefore, if a stateful LSTM is being used, the states must be reset in both the situations; after evaluating the network on a validation dataset as well as after making predictions.

The samples within an epoch are shuffled by default. This serves as a good practice when Multilayer Perceptron neural networks are used. If the LSTM state across samples are being preserved, then the order of samples in the training dataset are important and they must be preserved. This is done by setting the shuffle argument in the fit() function to False as follows:

```
for i in range(200):
model.fit(X, y, epochs=1, shuffle=False, batch_in-
put_shape=(20, 10, 2))
model.reset_states()
```

The conclusive remarks can be as follows, a critical concept to differentiate between stateful and stateless LSTMs is “when internal state is reset”.

a. **Stateless**: In this mode the internal state is reset after each training batch or the input batch while making the predictions.
b. **Stateful**: Here the internal state reset mechanism is triggered only when the reset_state() function is called.

If this is the only difference, then there might be a possibility that a stateful LSTM can be simulated with a stateless LSTM using a large batch size. In the next section the results from the above mentioned LSTMs and RNNs will be compared.

8 Results and Conclusion

In this section the results are presented for

a. A Vanilla RNN fed with a single series
b. A Stateful LSTM
c. A Stateless LSTM with the same configuration.
d. A Stateless LSTM with shuffling during training.

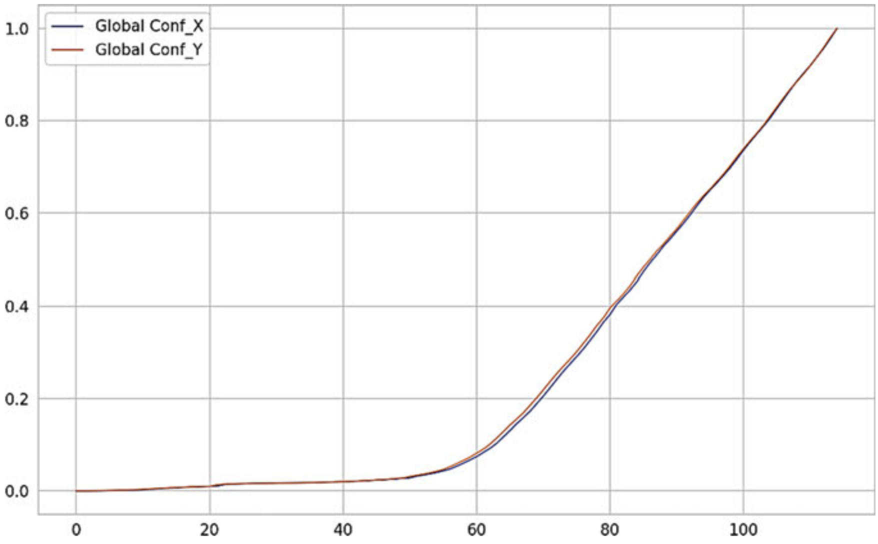


Fig. 8 Input univariate sequence data for global deaths due to COVID19

8.1 A Vanilla RNN

The vanilla RNN or conventional RNN was implemented from scratch, and the RNN was fed with a single sequence data of global death sequence.

The input possible is a single variate as the RNN is a vanilla one and little modifications can be done to make it suitable for multivariate data also. The input data is shown in Fig. 8, here the RNN will be trained for the Global deaths sequence data and it will be predicting the next number based on given input. Figure 9 shows the prediction made by RNN; it is quite deviated from the actual values as well as the trendline for the sequence.

8.2 Stateful LSTM

For the LSTM Implementation following parameters are fixed for the overall process

1. Number of LSTM cells—100
2. Number of Hidden Layers—01
3. Number of Training epochs—250
4. Optimizer—RMSProp
5. Loss Metric—Mean Absolute Error

The LSTM Model under Consideration is as follows:

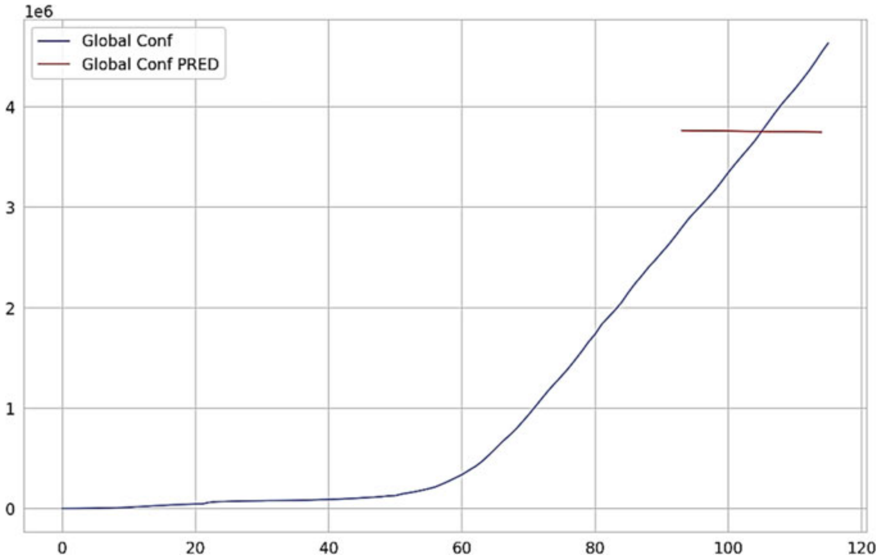


Fig. 9 Predictions by RNN for univariate sequence data for global deaths due to COVID19—shown in red markings over the actual values

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_5 (LSTM)                (5, 2, 100)               44000
_________________________________________________________________
lstm_6 (LSTM)                (5, 100)                  80400
_________________________________________________________________
dense_3 (Dense)              (5, 2)                    202
=================================================================
Total params: 124,602
Trainable params: 124,602
Non-trainable params: 0
```

The predictions by the Stateful LSTM is show in Fig. 10.

8.3 Stateless LSTM Without Shuffling

In this configuration, LSTM with same configuration is deployed but this time the model is stateless and the input data is shuffled. The results are shown in Fig. 11. The input patterns are shuffled for each batch or for an epoch, this practice is often followed to improve the Multi-layer Perceptron (MLP) network's generalizability during the training phase. A stateless LSTM does not shuffle the input data while training because the LSTM learns sequence of the patterns. The stateless LSTM with and without shuffling are implemented and the performance is compared here.

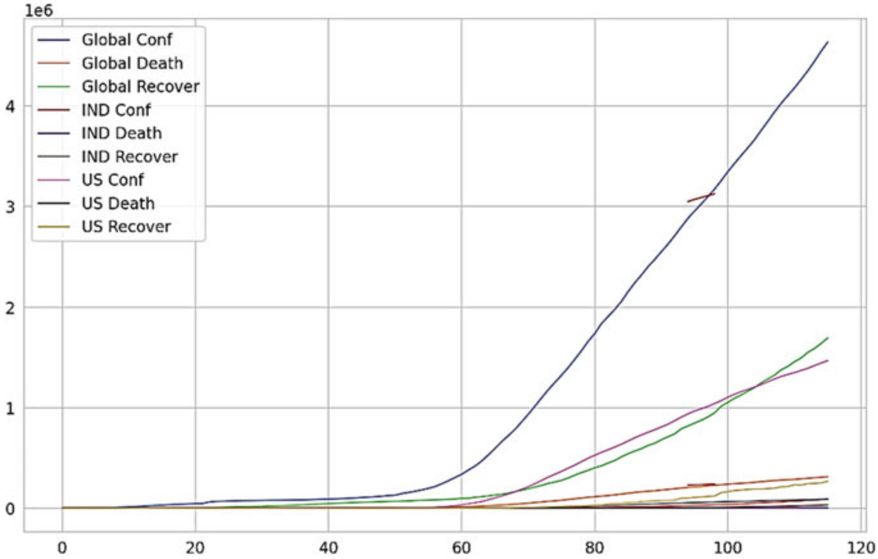


Fig. 10 Predictions by the **Stateful LSTM** for multivariate sequence data for global confirmed cases and global deaths due to COVID19—shown in red markings over the actual values

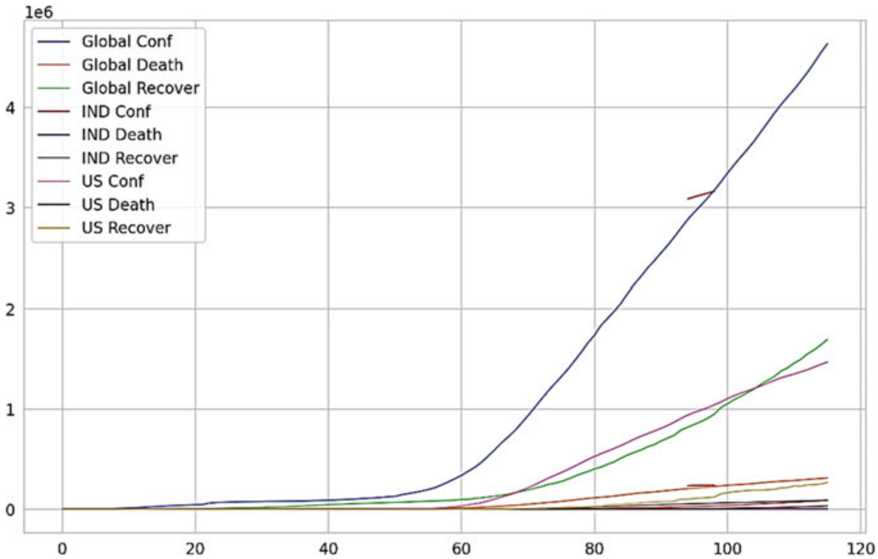


Fig. 11 Predictions by the **Stateless LSTM with shuffle = OFF** for multivariate sequence data for global confirmed cases and global deaths due to COVID19—shown in red markings over the actual values

The code changes to implement the stateless LSTM involve setting stateful = False in the LSTM layer and opting for the automated training epoch training instead of the manual one; further the shuffle flag is set to False.

8.4 Stateless with Shuffling

The stateless LSTM with shuffling of inputs are also implemented and their results are shown in Fig. 12. The results indicate that the stateless LSTM configurations outperform the stateful ones. The shuffling of training samples impacts significantly to the stateless LSTM; however, this heavily depends on the problem and training data.

The mean absolute error for the predictions made by each model discuss above is shown in Fig. 13. For the current problem of COVID-19 impact prediction; it is found that the Stateless LSTMs with Shuffling have given best performance. Vanilla RNNs also have a comparable performance.

One key observation is that these findings can further be made more robust by increasing the number of iterations for each experiment and further evaluation of the metrics to confirm the significance using statistical tests.

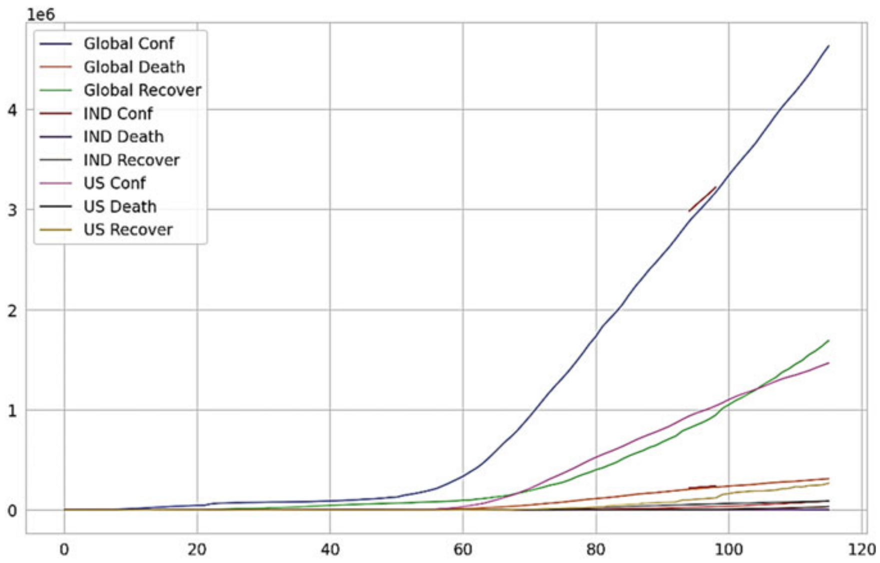


Fig. 12 Predictions by the **Stateless LSTM with shuffle = ON** for multivariate sequence data for global confirmed cases and global deaths due to COVID19—shown in red markings over the actual values

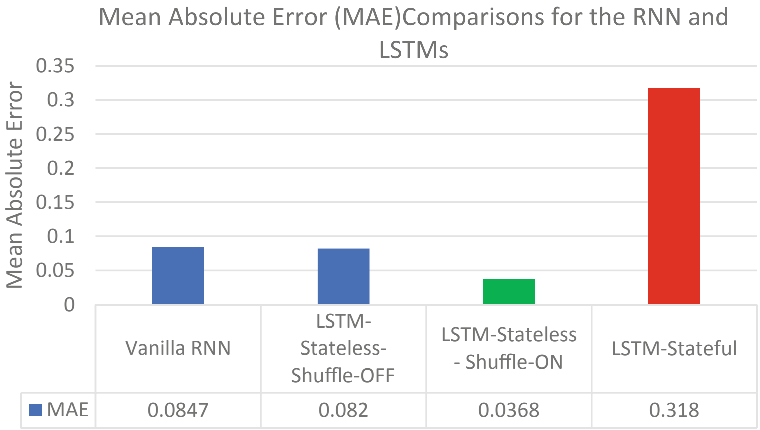


Fig. 13 Performance comparison for the RNN and LSTMs (green bar indicates the best performance and red bar indicates the worst performance)

Furthermore; these results apply to this specific problem of COVID-19 dataset prediction, the framing of the problem, and the selected LSTM configuration parameters including neurons, topology, layers, epochs and batch size. The dataset and the code for the above-mentioned LSTMs and RNN is available on Document Identifier DOI: 0.5281/zenodo.3952399 [19].

References

1. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2) (2019). https://doi.org/10.1109/TPAMI.2018.2798607
2. DiPietro, R., Hager, G.D.: Deep learning: RNNs and LSTM. In: Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 503–519. Elsevier (2019). https://doi.org/10.1016/B978-0-12-816176-0.00026-0
3. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings, pp. 6645–6649 (2013). http://arxiv.org/abs/1303.5778
4. Martens, J., Sutskever, I., Optimization, H., Martens, J., Sutskever, I.: Training deep and recurrent networks with. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7700 LECTU (2012). https://doi.org/10.1007/978-3-642-35289-8-27
5. Squartini, S., Hussain, A., Piazza, F.: Attempting to reduce the vanishing gradient effect through a novel recurrent multiscale architecture. Proc. Int. Joint Conf. Neural Netw. **4**, 2819–2824 (2003). https://doi.org/10.1109/ijcnn.2003.1224018
6. Hochreiter, S., Schmidhuber, J.: LSTM can solve hard long time lag problems. In: Proceedings of the 9th International Conference on Neural Information Processing Systems. NIPS'96: Proceedings of the 9th International Conference on Neural Information Processing Systems, pp. 473–479 (1996). https://doi.org/10.5555/2998981.2999048

7. Zeyer, A., Doetsch, P., Voigtlaender, P., Schluter, R., Ney, H.: A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings, pp. 2462–2466 (2017). https://doi.org/10.1109/ICASSP.2017.7952599
8. Kasongo, S.M., Sun, Y.: A Deep Gated Recurrent Unit based model for wireless intrusion detection system. ICT Express (2020). https://doi.org/10.1016/j.icte.2020.03.002
9. Bulín, M., Šmídl, L., Švec, J.: On using stateful LSTM networks for key-phrase detection. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11697 LNAI, pp. 287–298. Springer (2019). https://doi.org/10.1007/978-3-030-27947-9_2
10. Kondo, K., Ishikawa, A., Kimura, M.: Sequence to sequence with attention for influenza prevalence prediction using Google trends. In: ACM International Conference Proceeding Series, pp. 1–7 (2019). http://arxiv.org/abs/1907.02786
11. Zhu, X., Fu, B., Yang, Y., Ma, Y., Hao, J., Chen, S., Liu, S., Li, T., Liu, S., Guo, W., Liao, Z.: Attention-based recurrent neural network for influenza epidemic prediction. BMC Bioinform. **20**(S18), 575 (2019). https://doi.org/10.1186/s12859-019-3131-8
12. Pham, T., Tran, T., Phung, D., Venkatesh, S.: Predicting healthcare trajectories from medical records: a deep learning approach. J. Biomed. Inform. **69**, 218–229 (2017). https://doi.org/10.1016/j.jbi.2017.04.001
13. Chakraborty, R., Hasija, Y.: Predicting microRNA sequence using CNN and LSTM stacked in Seq2Seq architecture. IEEE/ACM Trans. Comput. Biol. Bioinf. 1–1 (2019). https://doi.org/10.1109/tcbb.2019.2936186
14. Hong, X., Lin, R., Yang, C., Zeng, N., Cai, C., Gou, J., Yang, J.: Predicting Alzheimer's disease using LSTM. IEEE Access **7**, 80893–80901 (2019). https://doi.org/10.1109/ACCESS.2019.2919385
15. Zhou, J., Lu, Q., Xu, R., Gui, L., Wang, H.: EL_LSTM: prediction of DNA-binding residue from protein sequence by combining long short-term memory and ensemble learning. IEEE/ACM Trans. Comput. Biol. Bioinf. **17**(1), 124–135 (2020). https://doi.org/10.1109/TCBB.2018.2858806
16. Li, H., Liu, Z., Ge, J.: Scientific research progress of COVID-19/SARS-CoV-2 in the first five months. J. Cellular Mol. Med. (Blackwell Publishing Inc.) (2020). https://doi.org/10.1111/jcmm.15364
17. Capelo, L.: Beginning Application Development with TensorFlow and Keras: Learn to Design, Develop, Train, and Deploy TensorFlow and Keras Models as Real-World Applications, 1st ed. Packt Publishing (2018). https://doi.org/10.5555/3265575
18. Rosebrock, A.: Keras vs. tf.keras: what's the difference in TensorFlow 2.0?—PyImageSearch (2019). https://www.pyimagesearch.com/2019/10/21/keras-vs-tf-keras-whats-the-difference-in-tensorflow-2-0/
19. Bharadi, V.A.: Analysis of Long Short Term Memory (LSTM) Networks in the Stateful and Stateless Mode for COVID-19 Impact Prediction (1.0). Zenodo (2020). https://doi.org/10.5281/ZENODO.3952399

Dr. Vinayak Ashok Bharadi received the B.E. degree in Electronics from Finolex Academy of Management and Technology, University of Mumbai, in 2002. M.E. degree in Electronics and Telecommunication in 2007 and Ph.D. in engineering in 2011 from NMIMS University, Mumbai, India. Presently he is working as Associate Professor with Department of Information Technology of Finolex Academy of Management and Technology He is life member of ISTE and area of interest are Deep Learning, Cloud computing, Big data analytics, Biometrics and DSP. He is a recognized Ph.D. Guide at Mumbai University. He has published more than 200 Research papers in various international Journals and Conferences.

Ms. Sujata S. Alegavi received the B.E. degree in Electronics & Telecommunication from IGCOE, University of Mumbai, in 2006. M.E. degree in Electronics and Telecommunication in 2010 and Currently pursuing Ph.D. in engineering since March 2015 at Mumbai University. Presently she is working as Assistant Professor with Department of Electronics Engineering in Thakur College of Engineering and Technology. Her area of research is Deep Learning, DSP and Image Processing.

Soft Computing Techniques for Energy Consumption and Resource Aware Allocation on Cloud: A Progress and Systematic Review

Sukhpreet Kaur, Yogesh Kumar, and Sushil Kumar

Abstract Cloud computing plays a vital role in storage and transfer of immense capacity data due to a rapid growth in size and the quantity of organizational tasks. There are many studies in which varied soft computing methods are applied to the cloud environment. In large data centers the cloud services indorse not only the energy consumption price of the substructure resources but also with a considerable growth in environmental costs. These subjects are significant requisites to decrease the energy cost and carbon footprint of cloud computing systems. To minimize energy consumption, the intelligent machines are required to achieve crossways numerous diverse machines, and strategies corresponding across the hardware and software layers to balance performance and energy, as well as to proficiently exploit multiple resources. Energy-efficient Cloud Organization Resource Allocation Framework is getting acceptance as it is paying operative consideration to cloud data management with an interpretation to achieve maximum revenue and minimum cost. The primary objective of the chapter is to conduct the systematic study and mapping of recent soft computing techniques to resolve the resource allocation and energy consumption problems in cloud computing. The chapter discuss the various soft computing techniques which are used in cloud environment for energy-resource allocation, workflow scheduling and performing the migration on cloud computing system. The first section of the chapter comprises of Introduction, motivation, background works which includes Framework for Energy and resource aware allocation using soft computing techniques, various issues, benefits of the work and application areas of soft computing techniques for cloud. The next section of the chapter highlighted the reported work which covers the detailed study of the researchers for energy efficiency and resource allocation using soft commuting techniques. The final section of the chapter discuss the comparative analysis which compares the work of

S. Kaur
Department of Computer Science & Engineering, Chandigarh Engineering College Landran, Mohali, India

Y. Kumar (✉)
Department of Computer Science & Engineering, Chandigarh Group of Colleges, Mohali, India

S. Kumar
Chandigarh Engineering College, Landran, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_9

different researchers by using various performance parameters such as execution time, power consumption, energy efficiency, resource utilization, response time and makespan.

Keywords Soft computing · Energy optimization · Resource allocation · Cloud computing · Workflow scheduling · Intelligent machines

1 Introduction

In this era of digital technology, the need of using the resources over the internet is of utmost importance. The solution of using the resources is Cloud Computing. It helps the users to store the data, access as well as process the data in a very effective manner. The users only need the connectivity with the Internet for accessing the data on cloud. Three basic components of any cloud are clients, data centers and distributed servers [1]. The clients are the primary component with whom all the end-users interact with for availing any services on the cloud. The data centers have the collection of many different types of servers deploying different applications. The concept of virtualization helps the data centers to work for clients from large distance. The distributed serves are available all over the Internet for hosting different applications. The cloud can be deployed as public, private, community and hybrid cloud. The services provided by the cloud includes Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). However, the effective usage of services on cloud requires the security, performance, load balancing, energy consumption and resource allocation to be addressed in a very effective manner [2].

The fundamental concept of cloud computing is to deliver the services related to IT to all the clients using minimal requirement of infrastructure. The cloud resources need to be managed properly for using them effectively. The cloud resources can be in form of hardware or software resources. For this, the Service Provider of the cloud provides these services in the virtual form and on the rental basis to all the users from existing pool of resources.

Load balancing is one of the important tasks while working on cloud as it directly impacts the performance of cloud. The resource allocation is performed so that no machine should be overloaded or under loaded, as this can lead to failure of the system. The different types of loads that are considered, includes load of the memory, various computations and load of the network. Resource allocation as well as load balancing will detect under loaded and overloaded machines and then balance the load among them while optimizing the different performance parameters. The balancing of the load on the cloud can be of two types—Static and Dynamic. If the environment is stable and homogeneous, then the static based load balancing algorithms can perform better. However, in case of heterogeneous and unstable networks, the dynamic based load balancing algorithms will perform better. Although, the dynamic based load balancing algorithms can work in both homogeneous and heterogeneous networks.

The energy consumption is the amount of energy consumed by all the devices in the network of the cloud. The overall consumption of the energy can be calculated by considering personal terminals like desktops, nodes used for networking like routers and switches and local servers like application servers. The energy can be conserved by either using energy-efficient hardware, energy aware scheduling technique or power minimization in clusters, wire and wired networks etc. The energy can also be conserved if the nodes in the network should be active and use power when they are performing any work, otherwise, they should remain inactive.

With the evolvement of the technology and techniques, the energy usage is also increasing at a great pace. The energy consumption is also increasing as the size of the network is increasing as more virtual machines are being deployed on cloud that in turn is increasing the physical machines for fulfilling the computing demands of the users. It is also one of the challenging issues, as the demand of the energy will increase in next 20 years as the data centers will expand.

2 Motivation

Numerous researchers worked very hard and solved many problems of cloud computing using the different techniques. However, still there is a chance of improvement in all the cases. The mapping of soft computing techniques was done in the different categories of cloud computing like in optimization of various tasks while their allocation and their performance, optimization of consumption of power during working on cloud, techniques for assuring security while working in cloud, selection of various services from the pool for the end-users as well as in optimizing the cost while using the different services on cloud [3].

The following points are described the motivation behind the use of soft computing techniques in energy efficiency and in resource allocation on cloud environment:

- Soft computing techniques also helped in managing the big data associated with the cloud. As cloud computing is one of the best choice for every user as it assures the maximum performance in reduced cost, so it need the intelligent solutions for sustaining the performance.
- As the handling of data of cloud is not possible manually, so for finding the intelligent solutions for coping with the features of cloud, soft computing was used.
- The various soft computing techniques used in cloud computing are fuzzy logic, Bio-inspired and Nature Inspired Algorithms which includes Genetic Algorithm, Swarm Intelligence like Particle Swarm Optimization, Ant Colony Optimization, Artificial Bee Colony Optimization, Bacteria Foraging Optimization etc. and the various hybrid systems that are developed using the combination the pros of these all methods.
- The various challenges in cloud computing can be effectively solved using the different techniques of soft computing.

- The energy consumption can be reduced by designing energy-aware policies as well as designing of special hardware and software for the cloud. Further, the energy can be used efficiently when the allocation of virtual machines will be performed in an optimal way by using optimization techniques.
- Using soft-computing methods, if the workload is well distributed among all the nodes, and then also the energy consumption is pretty less. It is also proved that if any node is being utilized optimally, then the energy consumed by those nodes is less as compared to the nodes, which are overloaded or under loaded.

3 Background

In this section, we have started with the introduction of the energy consumption as well as the resource allocation in the cloud computing using different soft computing techniques. Then the framework for the energy consumption is presented using the block diagram, different types of soft computing techniques that are being used till now are defined and the importance of soft computing techniques in both energy consumption as well as resource allocation is justified very clearly. Various research challenges and issues related to usage of soft computing techniques in the application of cloud is given. Various application areas of the soft computing techniques in cloud computing is also presented in this section. This section also has all the work reported so far in these two areas using any soft computing techniques by elaborating their all benefits as well as shortcomings.

3.1 Framework for Energy Consumption and Resource Aware Allocation

The computations that are being performed on the cloud have immense applications and also are one of the important parts of the today's digital world. Energy consumption is one of the challenging research issue as the data centers evolve with time in terms of complexity, size and power consumption. It needs to be performed in with utmost care as it is the future of supercomputing. So, there is a need that all the work that is to be performed on the cloud should be performed in less time as well as the consumption of the energy should be minimized while allocating all the resources to the jobs in effective manner. The scheduling of the jobs is also one of the important part along with the allocation of resources to them. The wrong scheduling can lead to wastage of large amount of energy.

In the given framework as shown in the Fig. 1, all the cloud users first send the request of jobs to the data centers. After that, the data centers with the help of scheduler will check and analyse the load using soft computing techniques. The load should be balanced while allocating the resources and the consumption of the energy should be minimized. Once it is decided that the jobs can be scheduled effectively

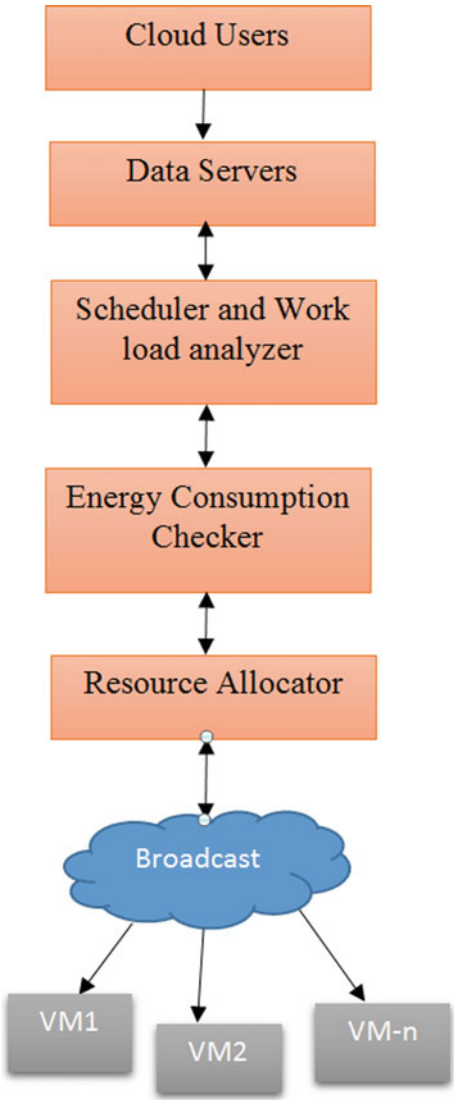


Fig. 1 General framework for energy-resource allocation on cloud

as all the resources are available for cloud users, then the information is broadcasted to all the virtual machines. The virtual machines can also be migrated in the cloud if the energy is consumed at the maximum level or there is some type of imbalance in the load allocated to the machines.

3.2 Types of Soft Computing Techniques

As discussed in previous sections, soft computing techniques are able to give effective solutions for many issues and challenges in cloud computing. The main feature of soft computing techniques are that they are able to solve the problems logically. As the cloud environment faces many issues due to involvement of number of different resources at different times as well as large number of users, so soft computing helped in providing the solutions for many of the problems of cloud. And also these solutions do not hamper the performance of the cloud.

Soft-computing techniques consists of algorithms based on either biological organisms or nature. As nature behaves in varying ways according to the situation as well as according to complexity of the problem, so this behaviour of nature can be used in real time problems to solve them [4, 5].

All the nature-inspired problems can be solved by using any one of the following categories:-

(i) All the computations that are executed for solving any problem are inspired by the nature.
(ii) By use of natural material for carrying computations.
(iii) By simulating the nature for performing computations.

The various bio-inspired soft computing techniques do not always give optimal solutions [6]. They are iterative in nature and provide optimal solutions for energy efficiency problems after doing allocation of virtual machines. The solutions given by soft computing is different from hard computing as they are intolerant to certainty and precision and have cheaper and more approximate solutions in comparison to them. The various techniques under soft computing are Fuzzy Logic, Neural Network, Hybrid techniques of both, various bio-inspired techniques like Genetic Algorithm, Particle Swarm Optimization, Ant Colony Optimization [7], Bacteria Foraging Techniques and number of other techniques based on the behaviour of the nature.

The different types of bio-inspired soft computing techniques along with the part of the nature they are based on is shown in the Table 1 as given below.

The basis of Soft computing techniques is Genetic Algorithm, which was developed by John Holland. This algorithm helps in finding the searching solutions iteratively using the process of natural genetics, which involves the process of mutation and crossover. These are under the category of Evolutionary Algorithms. In this, a population of constant size was taken and then using the generation of populations was declared using the iterative proves of inheritance and crossover. For each pair of individuals in a group, a fitness value is calculated through which they will be allowed to start a new generation. In a similar way, for each iteration, the best value of the fitness function is calculated and the solution is found in number of such iterations.

The operating rules of GA are as follows:

A. Selection—as per concept of survival for fittest, parents are selected.
B. Crossover—parents are open to mate, creation of next generation.

Table 1 List of bio-inspired algorithms

S. No.	Name of optimization algorithm	Based on
1	Genetic Algorithm [6]	Darwin evolution
2	Artificial Immune Systems [8]	Biological immune system
3	Ant Colony Optimization [7]	Ants foraging for food
4	Particle Swarm Optimization [9]	Flocking of migrating birds
5	Honey Bee Optimization [10]	Reproduction of Honey bee colony
6	Bacterial Foraging Optimization [11]	Optimal foraging of bacteria
7	Shuffled Frog Leaping Algorithm [12]	Frog population searching for food
8	Cat Swarm optimization [13]	Behavior of cats
9	Invasive weed optimization [14]	Colonized weeds
10	Monkey Search optimization [15]	Monkey in search of food
11	Water flow like algorithms [16]	Flow of water from higher to lower levels
12	Biogeography based optimization algorithm [6]	Organizations distributed over time and space
13	Fish School Search [17]	Oceanic fish
14	Cuckoo Optimization [18]	Cuckoo
15	Bat inspired algorithm [19]	Bats
16	Firefly Algorithm [7]	Fireflies and their flashing behavior
17	Dolphin Partner optimization [20]	Dolphins' behavior
18	Flower pollination algorithm [21]	Flowering plants and their pollination
19	Krill head optimization algorithm [22]	Krill individuals
20	Grey Wolf Optimization algorithm [23]	Behavior of wolves
21	Water Cycle Algorithm [16]	Rivers and streams and their flow
22	Social Spider optimization [24]	Spiders
23	Forest Optimization Algorithm [25]	Trees and their survival process and time

C. Mutation—making random changes in the parents to form children.

For the technique of Particle Swarm Optimization, that is simulating the social models of flocking of birds, schooling of fishes and swarming. It is heuristic model based on population modelling through simulation. It is proved to be better than other optimization techniques as Ant Colony Optimization (ACO) and Genetic Algorithm (GA) in terms of search space performance parameters and convergence rates. It is also proved to be best when thresholding is used in any process [26].

The basic idea behind PSO is that each particle is searching for its optimum value by moving. Since the particle is moving, hence it has its velocity and each particle remember its best position it reached during its journey, but that position cannot be best for all. For finding the best solution for all, the particles cooperate with each other by transferring the information about the best places they have visited so far.

This is done by the process of communicating the best fitness value of the neighbor of any particle to other particles.

3.3 Importance of Soft Computing Techniques for Energy Consumption

Energy consumption is one of the most important issue of cloud computing. It directly effects the performance of different algorithms in cloud computing. It is the one of the biggest challenge in case of cloud as different parameters need to be considered while dealing with soft computing algorithms. The saving of energy in cloud computing can help the service providers in number of ways. It helps in saving the overall cost as well as effective utilization of resources. The allocation of the tasks should be done by taking into consideration the real-time scenario of the systems in cloud. If we consider it, while allocation of tasks and it will increase the performance of the network by saving the energy consumed by the cloud. The maximum solutions given in cloud environment used the VM based allocations in cloud. In addition, this type of allocation helps in saving the cost of processing as well as the makespan timings of the network.

As the consumption of energy directly effects the Quality of Service (QoS) of the cloud, so the awareness for the saving of the energy is very important at this point of time. The users should also sustain the performance of the cloud at all levels of usage of cloud. It is quite difficult as the huge and large sized data centers need to be maintained by the cloud along with huge computations that are involved in this process.

In case of Soft Computing, energy consumption as well as work-loads for future can be predicted using fuzzy logic inference engine, neural network and linear filters. Energy consumption is the challenging issue as a large amount of cost was spent on data centers. The cooling gadgets used in servers have the biggest contribution in the consumption of the energy in case of cloud computing. Another major contributor are the multi-core processors, which are used in servers [5].

So, at the end it is concluded that, the energy awareness for its effective usage have one of the utmost concern in cloud computing. It also helps in providing the direct benefits to all the users as well as to providers. It is one of the current as well as the future concern for research as it may help in contributing to sustain the natural systems in the environment and helped in balancing the ecological system. It will also help in providing the benefits to the providers without compromising the existing services of quality.

3.4 Research Challenges or Issues

The various research challenges that are related to cloud computing can be started with handling of different types of uncertainties in the cloud environment. The various reasons of uncertainty in resources can be location of client, type of content to be shared on cloud, heterogeneous devices and different resource allocation services provided by client.

1. The main challenge in effective working of cloud environment is not having the adequate amount of resources that can be provided to all the users effectively on time. For example, the major resource can be bandwidth provided by the cloud for all the users that if not given to the users for using the multimedia content can affect the QoS of the cloud.
2. The prediction of the bandwidth requirement of all the users is also a big challenge. There are number of soft computing techniques that are applied for the prediction of resource requirement of the users so that they can be allocated effectively in advance.
3. There is also a problem of delay that is based on the distance between the clients and the data centers or servers that are providing cloud services. However, this challenge can be resolved much effectively as we can allocate the cloud according to the location of the users so that delay cannot effect much the performance of the system.
4. Another one of the biggest challenge of in this area is the balancing of load and handling of varying amounts of load by the cloud so that the overall performance of the system is not effected. This problem gave the researchers lot of work to do in this field. They gave number of solutions for solving the problem of load balancing but still there is a need of improvement in this. Load can be balanced by checking under loaded and overloaded machines in the cloud, so that the load can be shared and the overall performance will be optimized in case of dynamic workloads. No single model can be there that can handle all the cases, so it remains a challenge in the cloud environment. Load can also be balanced by the migration of the virtual machines between the data centers. Earlier, there was the technique in which the processes were migrated which was evolved from the process of virtualization. But now a days, the virtual machines can provide the best solutions for load balancing.
5. As discussed earlier, the management of usage of energy is also one of the biggest challenge in case of the cloud computing. In addition, according to researchers, the soft computing techniques can help in finding the appropriate solutions for saving the consumption of energy. The energy usage can be saved abruptly, if we can lower the energy used in powering up and cooling down of the data centers.

3.5 Application Areas

There are number of application areas of cloud computing in real time scenarios [27]. The various applications of energy effective soft computing techniques based cloud data centers can be summarized as below:-

1. The memory pre-fetchers can use the soft computing techniques for finding the best configuration. The variety of algorithms can be implemented which includes support vector machines, Naïve Bayes algorithm etc. [28].
2. Soft computing techniques can also be used for scheduling the jobs very effectively.
3. The time and other resources can also be predicted very effectively using the soft computing technique in cloud.
4. Various different types of soft computing techniques can also be used for predicting the number of requests that will help in deciding whether to increase or decrease the number of Virtual Machines in data centers.
5. The patterns of the usage of different resources by the cloud can also be predicted by the soft computing techniques.

4 Reported Work

This section contains the reported work that presents the findings in this field at its best. The work done so far on both energy consumption as well as on resource allocation has been shown along with their benefits and limitations. In this section, all the major work done by the researchers is shown for finding different types of research challenges and issues. The work done in cloud computing using the different soft computing techniques is shown using Swarm Intelligence (SI) into consideration. Swarm Intelligence (SI) has a major role in the energy consumption as well as resource allocation in case of cloud computing. SI is based on the natural behaviour of the swarms or insects like ants, bees, birds, fireflies etc.

4.1 Soft Computing Techniques for Energy Consumption

The consumption of energy is one of the biggest challenge in implementation of cloud computing. There are number of sources which consume energy and that are needed to be identified so that the consumption of energy can be minimized. The soft computing techniques can help in number of possible ways in designing an energy effective technique. **Wu et al.** proposed an energy—efficient technique for virtual machine placement using Genetic Algorithm. In this algorithm, the energy consumption by both physical machines as well as communication network was considered. **Ghafari et al.** also developed a soft computing technique that was based on ABC (Artificial Bee Colony) Optimization algorithm. This algorithm helped in conserving energy while considering utilization of CPU in the network cloud. **Hasan et al.** combined the advantages of Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) to develop a hybrid energy aware algorithm. **Reddy et al.** also developed an algorithm named as Modified Discrete Particle Swarm Optimization (MDPSO), which helped in achieving higher utilization of physical machines. This was achieved by decreasing the active machines in the cloud. This was done by finding the underutilized machines and sharing their load with other machines after balancing the load. **Duan et al.** added a novel technique, which consists of a prediction model and optimization model. The prediction model is based on fractal mathematics and optimization model is based on Ant Colony Optimization (ACO). **Kansal et al.** also proposed an algorithm for lowering the energy consumption by the nodes in cloud computing. This method used firefly optimization algorithm for the migration of virtual machines. In firefly algorithm, both local as well as global search is used for finding the optimal solution. The consumption of energy is saved by live migration of the virtual machines. **Meshkati et al.** also developed a hybrid approach of Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC) for reducing the execution time as well as consumption of energy by minimizing the cost function. As shown in the Table 2, the work has been compared by using different techniques, with their benefits and limitations.

4.2 Resource Allocation

There are number of resources present in data centers on the cloud that need to be allocated to the cloud users who requested for these resources. Resource allocation and load balancing is also one of the biggest challenge in the implementation of cloud computing as it directly effects the overall performance of the system. The use of different soft computing techniques helped the cloud users in allocating the resources in a much effective manner. **Yunhua Deng et al.** developed a dynamic load balancing method for virtual environments in which migration of load was done using selected cells and then global diffusion method was used for balancing of load. **Zhenzhong Zhang et al.** also proposed an algorithm for resource allocation in IaaS which predicted the requirements of resource of all virtual machines and then framed the load balancing algorithm. The work done was evaluated using CPU utilization and memory usage. **Dasgupta et al.** developed an optimized algorithm for load balancing based on Genetic Algorithm. This algorithm outperforms all the basic load balancing techniques using the same values for mutation and cross-over. **Mishra et al.** also proposed a technique for load balancing that was based in Ant Colony Optimization (ACO). This span of the cloud computing based services was minimized which made this method effective tool for balancing the load. **Babu et al.** proposed a resource allocation method that is based on the strategy of foraging of honey bees and this algorithm proved to be effective for heterogeneous clouds as well as independent tasks. **Wang et al.** also developed a novel technique based on job

Table 2 Summary of the work for energy consumption using soft computing

Author	Technique used	Benefits of the work	Limitation
Wu et al. [29]	Virtual machine placement using Genetic Algorithm	The method was able to tackle problems of different kinds as well as it can be scaled to solve the problems of larger size	The results have not considered accurate energy consumption models
Ghafari et al. [30]	Artificial Bee Colony Algorithm Minimal Migration Time (BEE-MMT)	It took into account utilization of CPU for conserving energy. The QoS metric used for this purpose is SLA Violation	The trade-off between energy consumption as well as SLA violation can be enhanced further
Hasan et al. [31]	Energy—Aware Hybrid Swarm Optimization Algorithm	Energy consumption was reduced by more than 42%. It also prioritized the allocation of the resources	The bottlenecks of wireless bandwidth and device capacity was not removed which were also consuming energy and increasing delay
Reddy et al. [32]	Modified Discrete Particle Swarm Optimization (MDPSO)	The proposed method saves the energy as well as minimizes the violation of service-level agreements	This method have not considered Input-output and CPU—intensive workloads while improving the energy consumption of the cloud
Duan et al. [33]	Fractal mathematics and Optimized Ant Colony Optimization (ACO)	In this method, energy was conserved as well as resource allocation was done effectively. The different demands of all the different components of the cloud environment was fulfilled	This algorithm cannot work with multiple types of resource scheduling scenarios that is heterogeneous systems
Kansal et al. [34]	Energy consumption by migrating virtual machines using Firefly optimization approach (FFO-EVMM)	Efficiency in usage of energy is achieved by optimal live migration of virtual machines This algorithm also works for heterogeneous systems that can be scaled to large size	This algorithm has not taken care of load balancing while utilizing the resources well
Meshkati et al. [35]	Energy consumption based on Particle Swarm Optimization and Artificial Bee Colony Optimization (HSF.ABC and PSO)	Hybrid approach helped in balancing both exploration as well as exploitation. Using this algorithm, both energy consumption as well as execution time was reduced	This algorithm cannot work for heterogeneous data centers. Moreover, the network and disk resources were not taken into account

(continued)

Table 2 (continued)

Author	Technique used	Benefits of the work	Limitation
Wen et al. [36]	Dynamic Integration method for energy efficiency virtual cloud data centers	This algorithm migrated the virtual machines in real live scenario that in turn helped in minimizing the consumption of energy	The impact of network resources is not added in the total consumption of the energy by the cloud

spanning tree and genetic algorithm named as JLGA helped in effectively allocating the resources and helps in balancing the entire system load very efficiently. **Joseph et al.** discovered a novel technique for resource allocation that was based on Family Gene approach that divided the processing among the different families which run in parallel for resource allocation. **Shojafar et al.** also developed a technique that was the hybrid of fuzzy logic as well as genetic algorithm (FUGE) and which helped in balancing the load by considering number of other static as well as dynamic parameters. **Priya et al.** also developed a fuzzy-based multidimensional approach for resource allocation as well as load balancing. Table 3 shows that summary of the work in the field resource allocation with their benefits and limitations.

5 Comparative Analysis of Soft Computing Techniques for Energy Consumption and Resource Aware Allocations

Cloud computing offers an vital part in storing and transferring of large capacity data due to a fast increase in size and the numeral of administrative accomplishments. Although number of soft computing methods are very prevalent and have been extensively functional to cloud environments, preliminary efforts to find appropriate methodical mapping studies in the field of soft computing methods in cloud computing environment. The comparative Table 4 shows that how soft computing techniques has been used in cloud environment for energy utilization and resource aware allocations. The table also presented the work has been done by different researcher and shows the outcomes in different fields such as purpose of work, type of soft computing technique used and reported output in terms of energy consumptions, response time, execution time and makespan.

6 Conclusion

In this chapter, we proposed an Energy-efficient Cloud Organization Resource Allocation Framework that has used different soft computing techniques for minimizing the consumption of energy as well as for the allocation of the resources to different cloud users. Since, the cloud computing is one of the emerging area in this era of digital technology, so the usage of soft computing technique proved to be effective in improving the overall performance of the cloud in terms of various quantitative parameters. This chapter includes the introduction to the cloud computing as well as to the various soft computing techniques. There are number of researchers who gave number of solutions for saving the energy usage in the cloud as well as allocation of the resources in the cloud in a very effective manner. However, the usage of soft computing techniques in the cloud always helped in improving the performance

Table 3 Summary of the work for resource allocation using soft computing

Author	Technique used	Benefits of the work	Limitation
Yunhua and Rynson [37]	Dynamic Load balancing for distributed virtual environments	The various benefits of work included the speed of load balancing in lesser elapse time and cost	This load balancing method was unable to maintain the effectiveness with changing topologies
Mondal et al. [38]	Optimized Stochastic Hill Climbing Method for balancing load	This method proved to give better results than FCFS and RR algorithms	This algorithm is centralized and have a single point of failure and also lacks in utilization of resources
Zhenzhong et al. (2013) [39]	Load Balancing Method in IaaS Cloud	This method helped in predicting and calculating the resource requirement of virtual machine and proposed a load balancing method	The network load as well as Disk I/O need to be added for increasing efficiency of work
Dasgupta et al. [40]	Load Balancing using Genetic Algorithm	This method outperformed all the methods like First Come First Serve (FCFS), Round Robin (RR) etc. by using fixed length strings for solutions	The method can be enhanced by doing the variation in mutation and cross-over
Mishra et al. [41]	Pheromone update method for Load Balancing using Ant Colony Optimization	This method helped in balancing the load for clouds of different sizes	This method does not consider fault tolerances
Babu et al. [42]	Honey-bee inspired load balancing algorithm	This algorithm also considered the priorities of tasks that were not taken care in heavily loaded virtual machines	This algorithm have not designed for workflows for dependent tasks
Wang et al. [43]	Job spanning tree and genetic algorithm	This method achieves effective load balancing using minimum make span. It also helped in scheduling of the tasks	The priority of the tasks were not considered. Moreover, the control strategy for the parameters of Genetic Algorithm should be global
Joseph et al. [44]	Resource Allocation based on Family Gene Approach	This method reduces the attempt of premature errors that occurred in case of Genetic Algorithm. Moreover, the mutation operator is dynamic	This method does not involve virtualization which is an important part of cloud computing
Shojafar et al. [45]	Load balancing using fuzzy theory and genetic algorithm	This technique allocates the resources to all the jobs based on processing speed, memory, bandwidth and job lengths	This method does not take into consideration the energy consumption while allocating the resources

(continued)

Table 3 (continued)

Author	Technique used	Benefits of the work	Limitation
Priya et al. [46]	Multidimensional Queuing Load Optimization Algorithm using Fuzzy Logic	This algorithm maximized the utilization of the resources in minimum processing time	The privacy of the data was not considered while sharing the resources within the cloud

Table 4 Soft computing techniques for energy consumption and resource aware allocations

Authors	Soft computing technique	Purpose of the work	Simulator	Reported output
Ejimogu and Basaran (2017)	Modified discrete particle swarm optimization (MDPSO)	The optimal strategy for energy aware virtual machine allocation which actually reduces the power consumption on the physical machines	CloudSim	Energy consumption = 2.25 kWh SLA violations = 11.26%
Mondal et al. (2012)	Stochastic hill climbing	Soft computing based hill climbing techniques has been used for the systematic manner to load distribution on cloud environment	CloudSim	Response time = 228.95 ms
Maryam et al. (2018)	Whale optimization algorithm	The task completion time and energy consumption problem has been resolved the WOA soft computing algorithm	MATLAB	Average time = 23.54 ms Average energy = 19.85
Pillai et al. (2017)	Genetic Algorithm	The presented study on GA-based method for improving the energy consumption and enhancement of multiprocessor systems using a weighted-sum approach	MATLAB	*E*max = 171.2 mJ Actual energy consumed = 151.1 mJ Percentage reduction = 11.75 mJ

(continued)

Table 4 (continued)

Authors	Soft computing technique	Purpose of the work	Simulator	Reported output
Rashidi and Sharifian (2017)	Fuzzy-based Decision Maker	To accommodate to exponential progress of requests, user requirements should be dispersed to diverse cloudlets and then clearly and dynamically conveyed to the servers allowing to the modern network and server status	Not Defined	Average = 0.5937 Standard deviations = 0.0312
Mishra et al. (2018)	Heuristic-based algorithms	Presented the approach which defined numerous load balancing methods in different (i.e., homogeneous, heterogeneous) cloud computing environments	CloudSim	Improved energy consumptions
Hasan et al. (2018)	Hybrid swarm optimization algorithm (HSO)	The reported work balancing the workload among the cloudlets using hybrid swarm optimization algorithms	CloudSim	Energy consumption = 0.0008 Execution time = 12.5–15.0 ms
Moganarangan et al. (2016)	Hybrid optimization algorithm	The presented study reduce the cost and computing infrastructure by energy scheduling using hybrid optimization technique	Simulation tool	Energy consumed (J) = 98.75 Energy improvement = 98.75%

(continued)

Table 4 (continued)

Authors	Soft computing technique	Purpose of the work	Simulator	Reported output
Jena (2017)	Task scheduling using clonal selection algorithm	The presented approach solves the task scheduling and energy consumption problems in cloud computing environment	CloudSim	Power consumption = 0.28–3.45 kWh
Fu et al. (2018)	Particle swarm optimization	The reported study reduce the energy consumptions of the cloud data centers	CloudSim	Resource utilization = 50%
Bergmann et al. (2013)	Simplified Swarm Optimization (SSO)	The presented approach reduce the energy consumption for distributed systems with dynamic voltage scaling	CloudSim	Makespan = 27.02% Energy = 79.21(J)
Milan et al. (2019)	Bacterial foraging optimization (BFO)	The presented study maps cloudlets to suitable virtual machines with minimizing the energy consumption and makespan also safeguarding the load balancing	CloudSim	Makespan = 105 ms Energy consumption = 600 kWh

parameters such as execution time, power consumption, energy efficiency, resource utilization, response time and makespan. So, it is concluded that if the nature inspired technique is implemented in cloud for providing various services, it will balance the load by migrating the virtual machines in real time scenario as well as the usage of energy will also be minimized.

References

1. Vouk, A.: Cloud computing-issues, research and implementations. In: Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces (2008)
2. Buyya, R., Yeo, C., Venugopal, S., Broberg, J., Brandic, I.: Cloud Computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Gener. Comput. Syst. **25**, 599–616 (2009)
3. Dillon, T., Wu, C., Chang, E.: Cloud computing: issues and challenges. In: 24th IEEE International Conference on Advanced Information Networking and Applications. IEEE Computer Society (2010)
4. Sukale, S., Biradar, D.: Review of nature inspired algorithms. Int. J. Comput. Appl. **109**(3), 6–8 (2015)
5. Kaur, Kumar, Y.: Swarm intelligence and its applications towards various computing: a systematic review. In: 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2020, pp. 57–62 (2020)
6. Binitha, S., Sathya, S.S.: A survey of bio inspired optimization algorithms. Int. J. Soft Comput. Eng. **2**(2), 137–151 (2012)
7. Dorigo, M., Maniezzo, V.: Colorni, Ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybern. B **26**, 29–41 (1996)
8. Dasgupta, D.: Advances in artificial immune systems. IEEE Comput. Intell. Mag. **1**(4), 40–49 (2006)
9. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the 1995 IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)
10. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Global Optim. **39**, 459–471 (2007)
11. Passino, K.M.: Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Syst. Mag. **22**(3), 52–67 (2002)
12. Eusuff, M., Lansey, K., Pasha, F.: Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization. Eng. Optim. **38**(2), 129–154 (2006)
13. Chu, S.C., Tsai, P., Pan, J.S.: Cat swarm optimization. In: Yang, Q., Webb, G. (eds.) PRICAI 2006: Trends in Artificial Intelligence. PRICAI 2006. Lecture Notes in Computer Science, vol. 4099. Springer, Berlin (2006)
14. Karimkashi, S., Kishk, A.A.: Invasive weed optimization and its features in electromagnetics. IEEE Trans. Antennas Propag. **58**(4), 1269–1278 (2010)
15. Zhao, R., Tang, W.: Monkey algorithm for global numerical applications. J. Uncertain Syst. **2**(5), 165–176 (2007)
16. Yang, F.C., Wang, Y.P.: Water flow-like algorithm for object grouping problems. J. Chin. Inst. Ind. Eng. **24**(6), 475–488 (2007)
17. Filho, C.J.A.B., de Lima Neto, F.B., Lins, A.J.C.C., Nascimento, A.I.S., Lima, M.P.: Fish school search. In: Chiong, R. (ed.) Nature-Inspired Algorithms for Optimisation. Studies in Computational Intelligence, vol. 193. Springer, Berlin (2009)
18. Rajabouin, R.: Cuckoo optimization algorithm. Appl. Soft Comput. (Elsevier) **11**(8), 5508–5518 (2011)

19. Yang, X.S.: A new metaheuristic bat-inspired algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) Nature Inspired Cooperative Strategies for Optimization (NICSO 2010). Studies in Computational Intelligence, vol. 284. Springer (2010)
20. Dhanya, D., Arivudainambi, D.: Dolphin partner optimization based secure and qualified virtual machine for resource allocation with streamline security analysis. Peer-to-Peer Netw. Appl. **12**, 1194–1213 (2019)
21. Yang, X.S.: Flower pollination algorithm for global optimization. In: Durand-Lose, J., Jonoska, N. (eds.) Unconventional Computation and Natural Computation. UCNC 2012. Lecture Notes in Computer Science, vol. 7445. Springer (2012)
22. Karthick, P.T., Palanisamy, C.: Optimized cluster head selection using krill herd algorithm for wireless sensor network. Automatika **60**(3), 340–348 (2019)
23. Kohli, M., Arora, S.: Chaotic grey wolf optimization algorithm for constrained optimization problems. J. Comput. Des. Eng. **5**(4), 458–472 (2018)
24. Mirjalili, S.Z., Saremi, S., Mirjalili, S.M.: Designing evolutionyary feedforward neural networks using social spider optimization algorithm. Neural Comput. Appl. **26**, 1919–1928 (2015)
25. Orujpour, M., Feizi-Derakhshi, M., Rahkar-Farshi, T.: Multi-modal forest optimization algorithm. Neural Comput. Appl. **32**, 6159–6173
26. Kumar, Y., Kaul, S., Sood, K.: A comprehensive view of different computing techniques—a systematic detailed literature review. In: International Conference on Advances in Engineering Science Management & Technology (ICAESMT) 2019, Uttaranchal University, Dehradun, India (2019)
27. Demirci, M.: A Survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments, pp. 1185–1190 (2015)
28. Liao, S.W., Hung, T.W., Nguyen, D., Chou, C., Tu, H., Zhou: Machine learning-based prefetch optimization for data center applications. In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, pp. 56–65. ACM (2009)
29. Wu, G., Tang, M., Tian, Y.C., Li, W.: Energy-efficient virtual machine placement in data centers by genetic algorithm. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) Neural Information Processing. ICONIP 2012. Lecture Notes in Computer Science, vol. 7665, pp. 315–323 (2012)
30. Ghafari, S.M., Fazeli, M., Patooghy, A., Rikhtechi, L.: Bee-MMT: a load balancing method for power consumption management in cloud computing. In: 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, pp. 76–80 (2013)
31. Hasan, A.R., Mohammed, A.M., Salih, Z., Ameedeen, M., Tapus, N., Mohammed, M.: HSO: a hybrid swarm optimization algorithm for re-ducing energy consumption in the cloudlets. TELKOMNIKA (Telecommun. Comput. Electron. Control) **16**(5), 2144–2154 (2018)
32. Dinesh Reddy, V., Gangadharan, G.R., Rao, G.S.V.R.K.: Energy-aware virtual machine allocation and selection in cloud data centers. Soft. Comput. **23**, 1917–1932 (2019)
33. Duan, H., Chen, C., Min, G., Wu, Y.: Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems. Future Gener. Comput. Syst. 154–166 (2016)
34. Kansal, N.J., Chana, I.: Energy-aware virtual machine migration for cloud computing—a firefly optimization approach. J. Grid. Comput. **14**, 327–345 (2016)
35. Meshkati, J., Safi-Esfahani, F.: Energy-aware resource utilization based on particle swarm optimization and artificial bee colony algorithms in cloud computing. J. Supercomput. **75**, 2455–2496 (2019)
36. Wen, Y., Li, Z., Jin, S., Lin, C., Liu, Z.: Energy-efficient virtual resource dynamic integration method in cloud computing. IEEE Access **5**, 12214–12223 (2017)
37. Yunhua, D., Rynson, W.H. L.: Heat diffusion based dynamic load balancing for distributed virtual environments. In: Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology. ACM (2010)
38. Mondal, B., Dasgupta, K., Dutta, P.: Load Balancing in Cloud Computing Using Stochastic Hill Climbing—A Soft Computing Approach. Science Direct C3IT (2012)
39. Zhenzhong, Z., Limin, X., Yuan, T., Tian, J., Shouxin, W., Hua, L.: A model based load balancing method in IaaS cloud. In: 42nd International Conference on Parallel Processing (2013)

40. Dasgupta, K., Mandal, B., Dutta, P., Mondal, J.K., Dam, S.: A genetic algorithm (GA) based load balancing strategy for cloud computing. In: Proceedings of Elsevier, Procedia Technology (2013)
41. Mishra, R., Jaiswal, A.: Ant colony optimization: a solution of load balancing in cloud. Int. J. Web Semant. Technol. **3**(2), 33–500 (2012)
42. Babu, D., VenkataKrishna, P.: Honey bee behavior inspired load balancing of tasks in cloud computing environments. Appl. Soft Comput. ASOC **1894**, 1–12 Elsevier B.V (2013)
43. Wang, T., Liu, Z., Chen, Y., Xu, Y., Dai, X.: Load balancing task scheduling based on genetic algorithm in cloud computing. In: IEEE 12th International Conference on Dependable Automaton Secure Computing, pp. 146–152 (2014)
44. Joseph, C.T., Chandrasekaran, K., Cyriac, R.: A novel family genetic approach for virtual machine allocation. Proc. Comput. Sci. **46**, 558–565 (2015)
45. Shojafar, M., Javanmardi, S., Abolfazli, S. (2015), "FUGE: A joint meta-heuristic approach to cloud job scheduling algorithm using fuzzy theory and a genetic method" Cluster Computing, Vol. 18, pp 829–844
46. Priya, V., Kumar, C.S., Kannan, R.: Resource scheduling algorithm with load balancing for cloud service provisioning. Appl. Soft Comput. J. **76**, 416–424 (2019)

Dr. Sukhpreet Kaur is working as Associate Professor in CSE department at Chandigarh Engineering College, Landran, Mohali. She has in total of 15 years of vast experience in teaching and research. She has done Ph.D. in CSE from I K Gujral Punjab Technical University, Jalandhar and has done her Masters in Technology in CSE from GNDEC, Ludhiana. The various research areas in which she worked includes Image Processing, Artificial Intelligence and Computer Vision.

Dr. Yogesh Kumar is working as Associate Professor at Chandigarh Engineering College, Landran, Mohali. He has done his Ph.D. CSE from Punjabi University, Patiala. Prior to this, he has done his M.Tech. CSE from Punjabi university, Patiala. His research areas include Natural Language Processing, Computer Vision, Machine Learning and Data Science.

Dr. Sushil Kumar is working as a Professor, Chandigarh Engineering College, Landran. He has in total of 15 years of vast experience in teaching and research. He has done Ph.D. in CSE from I K Gujral Punjab Technical University, Jalandhar and has done his Masters in Technology in CSE from GNDEC, Ludhiana. The various research areas in which he worked includes Computer Networks, IOT, Cloud Computing and Metaheuristics algorithms.

Automatic Segmentation and Classification of Brain Tumor from MR Images Using DWT-RBFNN

Hari Mohan Rai, Kalyan Chatterjee, and Anand Nayyar

Abstract Brain tumor segmentation and diagnosis is a very tedious and uncertain task for medical experts for its precise treatment. We have proposed an automatic segmentation and classification of tumor at from brain MR images using DWT-RBFNN classifier. The work is divided into two parts: segmentation and classification. The MR images are 1st preprocessed before applying segmentation and classification on it. In preprocessing stage, MRI images are denoised using hybrid technique DWT-ICA, and resized. The segmentation is performed using hybrid Ostu-canny edge technique and these segmented images are used for the classification of brain tumor. After segmentation stage, 13 types of feature are extracted using multiresolution DWT, these are Median, Variance, Mean, Standard deviation (SD), power spectral density (PSD), RMS (root mean square), Energy, entropy, correlation, homogeneity, skewness, contrast, smoothness. The dataset used in this work contains 443 MRI images of three categories, 105 benign, 223 normal and 105 malignant. The RBNN classifier is used on the segmented and feature extracted images for the classification of brain tumor into three classes, Normal, benign and malignant. To evaluate the performance of the classifier, seven types of evaluation metrics are used, accuracy, F1-score, specificity, classification error rate, recall, precision and overall accuracy. The proposed classifier performance was also compared with Feed forward neural network (FFNN) and Back propagation neural network (BPNN) classifiers. The accuracy result obtained using proposed classifier (RBFNN) is outstanding i.e.

H. M. Rai
Department of Electronics and Communication Engineering, Krishna Engineering College, Ghaziabad, UP, India

K. Chatterjee
Department of Electrical Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India
e-mail: kalyanchatterjee@iitism.ac.in

A. Nayyar (✉)
Graduate School, Duy Tan University, Da Nang 550000, Vietnam
e-mail: anandnayyar@duytan.edu.vn

Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_10

100% on the test dataset as compared to FFNN and BPNN 95.92% and 97.96% respectively.

Keywords Brain tumor · Magnetic resonance imaging · Neural network · Image processing · Biomedical imaging · Discrete wavelet transform · Brain tumor detection

1 Introduction

One-tenth people who have significant symptoms, such as tumors or cancer or infections, are diagnosed too late or incorrectly, and 53.9% of them are either completely disabled or died due to medical error According to the research report [13]. The death rate are very high due to the misdiagnosis in the case of cancerous or tumors, which makes it most dangerous and fatal diseases for humans, it must be detected, diagnosed and cured in early stages. It occurs in almost every part of the human body but the human brain is the part where the tumor is the most. Brain tumor is life-threatening, if it is a cancerous, then utmost survival rates after detection of this type of brain tumor are about 9–12 months. By 2018, about 7 lakh people in the US are suffering from brain tumors and 40 to 50 thousand cases are being reported annually in India [34, 45]. About 30% of them suffer from secondary tumors (malignant) and 70% from primary tumors (benign), and only in the USA will be diagnosed with approximately 87 thousand primary brain tumor cases by 2020 [34].

Therefore we can say that the medical and healthcare sectors are more prone to human life, automated decision making in the medical field requires much precision and efficiency. Hence the accurate, automatic and early detection and diagnosis of critical diseases such as cancerous, heart diseases, and vital infections may saves the millions of lives around the globe. The death rate due to brain tumor is increasing very rapidly due to non-detection or erroneous detection manually in recent few years [41]. Hence early segmentation, recognition, and classification of brain tumor is very much important to save patients' lives and also to help medical experts in the treatment. Not only its segmentation and diagnosis but also its automatic identification with 100% accuracy is required to cure it at initial stage. Magnetic resonance imaging (MRI) images are the most important source to diagnose the brain tumor because of its unique characteristics compared to other imaging techniques.

Any kind of irregular, uncontrolled, abnormal, and unwanted growth of the tissue may cause harm to the neighbors healthy tissue is called tumor [21]. It has many categories but for the easy understanding and based on its cancerous nature it is divided into two types, initial and non-cancerous tumor is called benign and cancerous and final stage tumors are known as malignant. Also the diagnosis and treatment of the tumor is based on its type, that means the early or first stage tumor recognition (benign) is very imperative [41].

There is a grading system available for the tumors and its types which vary from grade 1 to grade 4, according to (WHO) world health organization. These grades are being used for categorizing the various types of tumor, where Initial 1st and 2nd grades are termed as low-level grades that identify the benign type of tumor. High-level grades, 3rd and 4th grade are at the fast-growing stage which identifies the malignant tumor [46].

A benign tumor means the slowly growing controlled growth of tissue, because of a certain reason, it is not cancerous. The abnormal, rapid and uncontrolled growth of cells which is cancerous is known as malignant. In the medical field, the biopsy procedure is used by the doctors to detect tumors whether it is benign or malignant [18, 29].

A benign tumor is not much worrying unless it is affecting nearby tissues, blood vessels, or nerves by putting pressure on it which may cause damage. Benign tumors do not spread but it should be removed by surgery, once it is removed do not appear in the same place. This type of tumor may grow large sometimes, usually, it is not dangerous but it can press blood vessels or may block channels [12]. It becomes more dangerous when it grows in the human brain because it affects the normal brain tissue by applying pressure on it. Few types of benign tumors such as intestinal polyps may become a malignant tumor, that's why such types of tumors are categorized as pre censorious and must be removed before reaching the censorious stage [7, 31]. Malignant tumors are very dangerous and it must be treated as early it is detected because it not only spread to nearby tissues but also cancerous. It is made up of cancer cells, few cancer cells may pass into lymph nodes or bloodstreams where it may spread into adjacent tissues within the human body [6, 46].

The structure of this chapter is arranged as; Sect. 2 presents the literature review carried out on the brain tumor segmentation and detection. Section 3 elaborate the materials and methods employed in this chapter including MRI database and WT. Section 4 explains about the preprocessing steps applied for filtering the MR images and tumor segmentation technique utilized in this chapter. Sections 5 describes about the proposed classifier applied for classification of tumors. Section 6 details the results obtained through the experiment and elaborates the analysis and discussion about the results, and the final Sect. 7 provides concluding remarks and future work of this chapter.

2 Literature Review

The machine learning (ML), Internet of Things (IOT), and several analysis techniques assist health expert in diagnosing the health conditions of the human beings. The many researchers have contributed significantly in this area such as Internet of medical things (IoMT) [32], healthcare 4.0 [25], Internet of Nano things (IoNT) [36], Swarm intelligence in healthcare [17], Healthcare models and algorithms [27], cancer diagnosis using Evolutionary Algorithm (EA) [28].

The medical images are commonly used in the diagnosis and treatment of the ecmost human diseases. Among them, Medical resonance Imaging (MRI) images are most preferred for tumor segmentation and detection and it contains maximum pixel information. In the early stage, cancer has not spread much but in the later stage, it has a chance of spreading in more parts of the body. There are several research and techniques have been proposed by many researchers, scientists and academicians for the automatic detection of brain tumor. Also many computer based algorithms and techniques have been proposed in recent decade using MRI images such as, ELM (Extreme Learning Machine) [30], RGBS (region-growing based segmentation) [42], ANN (Artificial Neural Network) [33], SVM (Support Vector Machine) [5], SOM (Self Organizing Map) [48], Competitive PNN [50], DWT-Probabilistic Neural Network (PNN) [46], CFT (Clustering Fusion Technique) [4], Ensemble Classifier [14], FCMC (Fuzzy C-means Means Clustering) [40], and many more.

Thapaliya et al. [43] projected a process of segmentation of brain tumor from magnetic resonance (MR) images. The proposed algorithm was validated on 8 patients' data and obtained a Jaccard index value between 83.49 and 93.76% and Dice coefficient values between 81.13 and 95.42%. Nachimuthu and Baladhandapani [30] presented the method to enhance the accuracy and efficiency of the classifier, by combining MRI and MR spectroscopy based on multi-dimensional co-occurrence matrix, to detect the normal, pathological and fluid tissues from brain magnetic resonance images. ANN method to classify the brain tumor in two types benign and malignant was proposed by Nazir et al. [33]. The classification of images in two categories were done by ANN classifier with the accuracy of 91.8%. Ji et al. [20] proposed MRI image segmentation technique using a generative asymmetric Gaussian mixture model (GAGMM) based on spatial constraint, where by including the spatial information, the asymmetric distribution can easily be modified. Based on the prior and posterior probabilities of the clusters, their asymmetric model was developed.

Ilhan and Ilhan [16] discussed a method for differentiating the cancerous and non-cancerous tissue from brain MR images. They used a new thresholding technique to segment the tumor portion that could clearly characterize the tumor and would aid the medical practitioner in diagnosis. The Algorithm was validated on 100 MR images and provided better results as compared to other techniques. TBRO (Threshold Based Region Optimization) method for the segmentation of brain MR images was proposed by Kanmani and Marikkannu [23]. They performed the experimentation and achieved very high rate of accuracy of classification 96.57% into normal and abnormal class. Ahmed et al. [2] performed experiment on MR Dataset to classify into normal and abnormal class. They recommended an ANN model with uniting GW (gray wolf) optimizer and acquired 98.91% of accuracy. Jeevakala and Brintha Therese [19] proposed an enhancement method for sharpening the images using LP (Laplacian pyramid) and SVD (singular value decomposition) to enhance the perceptibility and segmentation of MR images. The sharpening enhancement method for denoising the MR images along with sharpening the edges. They used performance metrics AMBE (absolute mean brightness error), MSSIM (mean structure similarity index) and PSNR (Peak signal to noise ration) for measuring the performance of the offered method. Devkota et al. [11] suggested a MMR (Mathematical Morphological Reconstruction) approach for diagnosis and detection of brain tumor at its early stage. They denoised and removed the artefacts from the MR images in preprocessing stage and then segmented it. Based on the statistical and texture based feature they classified the tumor into two classes benign and malignant.

Deepa and Sam Emmanuel [10] proposed a fused feature based adaptive firefly backpropagation neural network (AFBPNN) for classification of tumors which includes preprocessing, extraction, selection, and fusion of features to improve the efficiency of classification. The KPCA (kernel principal component analysis) method was employed for selecting the features and GRBF (Gaussian radial basis function) technique was utilized for feature fusion purpose. Kalaiselvi et al. [22] proposed novel method to generate RGB images by unification of the color channel formed from multiplexed or multimodel images. They segmented the tumor of brain from MRI-BRATS dataset used in the experimentation. Virupakshappa and Amarapur [47] advised a new method for feature extraction which operates on multi-level WT decomposition and chief descriptors such as Gabor, GLCM, and moment invariants features were extracted for wavelet coefficients. The Whale optimization algorithm (WOA) based adaptive artificial neural network (AANN) classifier was used in their work for classification purposes. The simulation was performed on the MATLAB platform and AANN classifier provides accuracy of 98% as compared to other classifiers.

Regardless of abundant contributions, the automated MRI tumor segmentation and its classification is still a challenging task for the researchers. Although, various techniques and algorithms have been proposed and implemented for segmenting the tumor from brain and its classification but still the performance accuracy has scope to improve. Only some researchers have applied segmentation and classification together for enhancing the performance accuracy of classifier. The automatic classification of tumors at early stage (benign) have not explored by most of the researchers.

To provide the solutions to the above concern, this chapter proposed the auto-segmentation as well as classification of brain tumors at early stage (benign). The brain MR images are preprocessed and segmented then it has been classified into three categories normal (without tumor) benign (non-cancerous), and malignant (cancerous) based on morphological and statistical features using a radial basis function network (RBFN) with 100% accuracy on test dataset. Even though the brain MR image datasets are not available in abundant nature, especially early stage tumor images (benign), we have used 443 brain MR images for automatic segmentation and predication of tumor.

3 Materials and Methods

The proposed method for automatic segmentation and classification of brain tumor utilized for this work is demonstrated in Fig. 1. The methodology of our work is mainly divided into five parts, preprocessing, segmentation, feature extraction, post processing, and classification. This section will details about the method and materials used for the experimentation purpose.

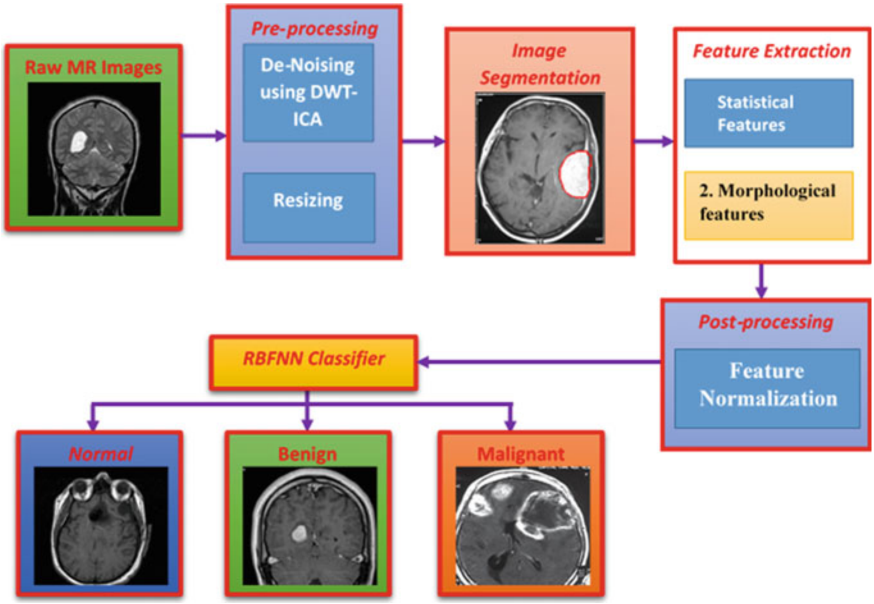


Fig. 1 Proposed algorithm for tumor segmentation and classification

3.1 Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging, mostly abbreviated as MRI, is a type of bio-imaging technique used to diagnose certain types of diseases in humans. It is one of the most imperative bio-imaging among all types (X-ray, CT, PET, ultrasound, SPET, etc.), specifically to detect tumors or cancers from the human body from the early 80s. The magnetic resonance scheme, which uses magnetic and radio waves, is used to produce MRI images of different parts of the body. The entire image of the body is scanned based on technology where a magnetic field is generated inside the human body and the atoms of the body are affected by the magnetic field and this increases the energy level in that area. In the deficiency of field, the human body relaxed the atom and sends radio magnetic signals by itself, so the scanner gathers these waves and form them as an image. The human body comprises of approximately 70% water and the constituent of water are hydrogen and oxygen, so hydrogen atomic nuclei play a significant role in the scanning of MRI images. The level of hydrogen present in our body decides the contrast (brightness and darkness) of the MRI images. This means that the area where the hydrogen content is higher will appear brighter, such as adipose tissue and where the hydrogen level is lower, such as in bones that it will appear darker, in MRI scanning.

The human body is scanned by an MRI scanner, along with a high pixel and clear picture of bones, which is why it is mostly used to examine the spinal cord and brain [39]. In MRI scanning, MR images are categorized in mainly three types T1, T2, and

PD weighted. T1-weighted MRI shows exceptional anatomic feature details but has the disadvantage that it exerts the contrast between abnormal and normal tissue. While T2-weighted images do not offer a decent description of anatomy, it does provide an excellent contrast between normal and abnormal tissues. The PD weighted have an only variation of wave from the modification of the number of spins, and it generates the contrast by reducing the impression of T1 and T2 weighted images [9].

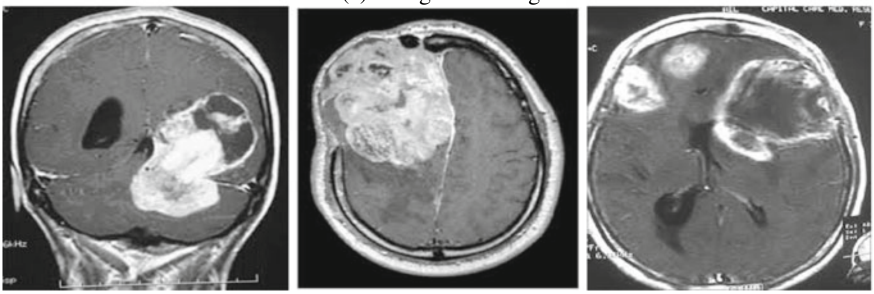
3.2 MRI Database

The dataset also have very useful impact in image processing and pattern recognition problem. We have collected 443 MRI images from the various sources including normal MR images (without tumor), and tumorous images (benign and malignant). The dataset contains 105 number of benign tumor (non-cancerous or early stage), 105 malignant (cancerous or final stage) tumor and 223 images without tumor. Images are of various shape and size that means heterogeneous in nature and with 3 channel (RGB) images. Dataset consists of brain MRI images having benign, malignant tumors and normal images which was downloaded from the Harvard Medical School website, ADNI dataset and OASIS dataset. We have used T2 weighted MR images, because T2 weighted images have a clear vision and higher contrast as compared to T1 weighted images [1, 8, 26, 35]. The sample test images used for the segmentation of brain tumors are shown in Fig. 2.

3.3 Wavelet Transform (WT)

To represent, characterize, and analyze the multi-resolution images, the wavelets are the most suitable and general tool. In simple word wavelet is also known as “Short or small wave”. For some applications it is required to analyze time and frequency together. STFT is also used for short and small wave and only to explore very short portion of the signal at a particular time, signal windowing method is used. It is quiet similar of wavelet but it has fixed and stable window size, also its window size is very narrow which results very reduced resolution in frequency. To overcome the hitches of STFT especially in frequency resolution wavelet transform is used. The comparison based on FR (frequency resolution and TR (Time resolution) between STFT and WT is shown in Fig. 3. To exemplify and characterize the signals of diverse frequencies of diverse resolution the MRA (Mult-Resolution Analysis) are applied. In MRWTA (Multi-Resolution Wavelet Transform Analysis), we mainly achieve high resolution in time and low resolution of frequency at upper frequency level similarly at lower level of frequency time resolution degrades and resolution of frequency becomes better. Hence, MRWTA is best suitable for components of high frequency with smaller time duration and components lesser frequency at lengthier time duration.

(a) Benign MR images



(b) Malignant MR images

Fig. 2 Test images. **a** Benign tumors, **b** malignant tumors

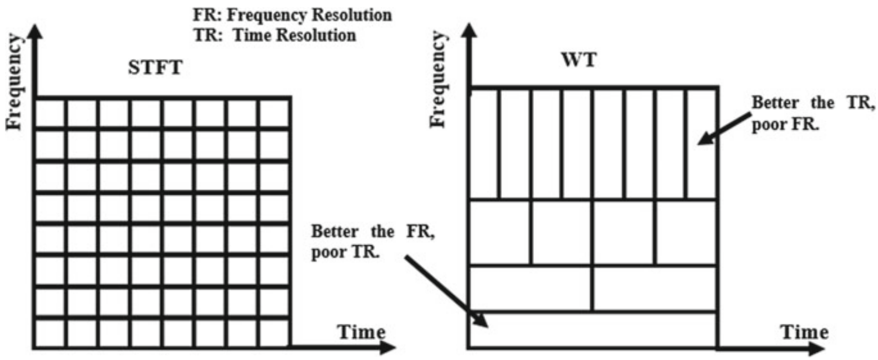


Fig. 3 STFT versus WT in terms of resolution of time and frequency

WT is one of the most dominant and powerful tool for characterizing and analyzing the signals in terms of time domain–frequency domain. As we know that the any types of images are nothing but 2D or 3D signals, hence MRDWTA techniques has been made use for preprocessing of MR images. Wavelet is basically also an analysis tool which perform based on the scale (time–frequency) i.e. it has finite length of windowing function. It includes some functions that satisfy certain mathematical constraints which are involved to demonstrating the signals data or some additional functions. The WT adopts a prototype function, termed as mother wavelet and also this porotype is used for producing additional WT functions. All generated WT functions are the shifted or translated and compressed or dilated version of original mother wavelet. Wavelet is mainly divided into two parts; CW (continuous wavelet) and DW (discrete wavelet).

The CW is explicated by the shifted and compressed function of mother or prototype wavelet ø(t) in Eq. (1). The CW transform in time and frequency (FT) domain can be expressed as:

$$\emptyset_{xy}(t)\frac{1}{\sqrt{x}}\emptyset\left(\frac{t-y}{x}\right) \leftrightarrow \emptyset_{xy}(\mho) = \sqrt{x}\emptyset(a\mho)e^{-iy\mho} \quad (1)$$

In Eq. (1), ‘x’ and ‘y’ are used as the random real numbers which denote shifting (dilation) and compressed (translation) parameters correspondingly [3].

Prototype wavelet for CW is expressed as:

$$\emptyset(t) = \emptyset_{1,0}(\mathrm{t}) \quad (2)$$

The shifted version of wavelet where x $\neq$ 0 and y $=$ 0, expressed as:

$$\emptyset_{x,0}(t) = \frac{1}{\sqrt{x}}\emptyset\left(\frac{t}{y}\right) \quad (3)$$

From the Eqs. (2) and (3), $\emptyset_{x,0}(t)$ is scaled wavelet by “x” in terms of time and also its amplitude is scaled by “1/x” with respect to the MW (Mother Wavelet) $\emptyset(t)$. Also, the signal represented with the above function can be stretched or expanded and compressed or contacted with respect to its MW along time axis by changing the “x” value. When the value of of “x” will be greater than 1 then it will expand and if its value is less than 1 then it will compress the MW along time axis, hence “x” is termed as scaling factor of the MW. Similarly, the value of “y” will shift in left direction or right direction with respect to time axis, hence it is named shifting factor and translation factor in terms of frequency and time respectively. The MW will shift to the right when the “y” value is kept above zero, and when its value is less than zero, the prototype wavelet shifts to the left of the time axis.

3.4 DWT Image Filtering

The MRDWTA (Multi-Resolution Discrete Wavelet Transform Analysis) is used for the filtering of raw MR images. The reason behind using DW transform over CW transform is that its computation efficiency is so efficient that it is also known as fast WT (FWT). And also the CW transform has few weakness such as it is impractical and redundant in nature. The DW transform overcome both negative aspects of CW transform very efficiently because of discrete in nature and fast processing.

The technique of filter bank theory i.e. down-sampling filter bank along with QMFs (Quadrature Mirror Filters), are used in DW transform. This means that the DWT has a set of digital filter banks made up of LP (low pass) and HP (high pass) filters. Low pass filtering gives an estimate of a signal (approximation), while high pass filtering gives a description that constitutes the difference between two successive estimates (details).

Since images are generally 2D signals hence for its processing the two-dimensional WT are used, where 2D-DW transform also termed as 'multidimensional' WT.

The DWT expressed in terms of MW and its scaling factor:

$$g(n) = \sum_{k} c(k)\Psi_k(n) + \sum_{i=o}^{\infty} \sum_{k=-\infty}^{\infty} d(i,k)\emptyset_{i.k}(n) \tag{4}$$

In Eq. (4), ø(n) is the MW and Ψ(n) is the scaling factor. MW as well as scaling factor are contracted by a factor 2 along the time axis as "j" is incremented by one. The index k represents the translations of MW.

The scaling factor and MW must satisfy a pair of recursive equations to form a basis for representing the signals:

$$\emptyset(n) = \sum_{l} w(l)\sqrt{2\emptyset(2n-l)} \tag{5}$$

$$\Psi(n) = \sum_{l} s(l)\sqrt{2\emptyset(2\mathrm{n}-1)} \tag{6}$$

The co-efficient of scaling factor is given by $s(l)$ in Eq. (5) an MW coefficients is $w(l)$ in Eq. (6).

4 Preprocessing of MRI Data

Preprocessing of image records is the most important part in the analysis of images and anchors, due to the fact that inaccurate information with unwanted data can lead to poor performance of specific broad classifications. The MRI image dataset used for the experimentation is not homogeneous, it have different shape (width × height) and of depth (channel) 3 because of RGB or color images. Also these MRI images are may be corrupted with different kinds of noises so, it is required to remove the noise. In the preprocessing stage, we will 1st resize the images of same width and height (256 × 256) and we will denoised it by applying hybrid denoising method using DWT-ICA.

4.1 Resizing of MRI Images

The dataset employed for the experimentation are heterogeneous in nature, hence it is suggested to resize it of equal width and height that means in square shape. The reason behind the selection of square shaped images i.e. the same height and width as almost all algorithms for image analysis is designed only for the image of the square shape otherwise we have to apply some different techniques to process it. Also it is desirable to resize the images in such a way that it could maintain all its information without losing or degrading its quality. For the experimentation work, we have resized the images of height 256 and width 256 to maintain the shape of (256 × 256).

We have converted the images into gray scale for the segmentation step to apply the segmentation algorithm efficiently. A grayscale image is a range of shades of gray without apparent color. The range of image pixels varies from 0 to 255 which is a total of 256 pixels, where the minimum value 0 means complete black color and the maximum value 255 means complete white color. To segment the images, it is required to convert it to the binary color means only true and false for that the gray scale image conversion is required.

4.2 Image Denoising Using MRDWT-ICA

Since images are collected from various sources that are traditionally available for downloading, there is a possibility of noise present in it. The 2nd step in the preprocessing is to denoised or filter it before applying segmentation algorithm into it. The filtering of images are performed using hybrid algorithm of denoising DWT-ICA, and detail steps for denoising of medical (MR) images is described in Rai and Chatterjee [37]. The denoising using DWT-ICA is briefly detailed in the following steps.

(1) The resized images of shape 256 × 256 are 1st normalized by applying centering operation and it is done by deducting mean vector (M) = E(x), hence the image will have zero mean variable.
(2) Using singular value decomposition (SVD), the whitening is competed which provides transformation matrix in the 2nd step.
(3) Fast-ICA is used in 3rd step to determine the de-mixing matrix ‘ Ꞓↄ’.
(4) In 4th step, N × 1 column vector matrix is randomly initiated and named ‘W’.
(5) A new function was introduced in the 5th step which is given by:

$$\hat{\mathrm{W}}(\mathrm{n})^{+} = E\{Xg(\text{Ꞓↄ}^{T}X)\}\text{-}E\{g_{0}(\text{Ꞓↄ}^{T}X)\text{Ꞓↄ}\}$$

Here “W” and W(n) + are known as temporarily variable.

(6) The new weight matrix is defined in the 6th step, given by $\hat{\mathrm{W}}(\mathrm{n}) = \frac{\hat{\mathrm{W}}^{+}}{\hat{\mathrm{W}}^{+}}$ here 'n' is the number of iterations.

(7) The conversion of data is checked at 7th step. If it is not converged then return to step 5.

(8) De-mixing ω matrix from step 3 is maximized by maximizing the source "S" which is the non-gaussianity measurement. Then the image with noise source is generated using the equation:

$$\hat{S}_{Noisy} = \omega \times Z \tag{7}$$

Equation 7, provides the IC approximation of noisy image source. The denoised IC estimate will be obtained by applying multiresolution DWT.

(9) The multiresolution DWT is applied on the noisy IC estimates obtained from Eq. 7. The DWT decomposition is applied using 3rd coefficient of daubechies (Db3) up to 5th decomposition level. Then VisuShrink thresholding is applied $V = \sigma\sqrt{2\log N}$, where level of noise variance is "σ" and pixel number of images is given by "N".

(10) Then the images are reconstructed and also mixing matrix named "A" is obtained by applying the ω

(11) At last step, the denoised images (MR) is obtained from Eq. (8).

$$\tilde{X}_{Denoised} = A \times \tilde{S}_{Denoised} \tag{8}$$

4.3 Tumor Segmentation

The segmentation is the process of splitting or separating any digital image into several slices (segments) is called segmentation. The purpose of image segmentation is to present it in the simplest form and to make it very informative to analyze. There are several methods for segmenting the images where threshold segmentation is one of the simplest and efficient method. In this method the pixels of the image are separated by contrast of the image, mainly on the basis of intensity and then the specific area of the image is segmented based on the application. Basically there are two methods of thresholding based on the image characteristics, if the threshold only reliant on the gray-level of the image and also the threshold value only correlated to the features of the image pixels, then it is called global thresholding. Local thresholding is defined because the threshold value depends on two threshold values. In this technique the raw input images are divided into several smaller sub regions and different threshold values are selected for each region.

4.4 Otsu's Thresholding

It is one of the method of global thresholding, where threshold value only depends upon the gray pixels of the images. The Otsu thresholding method is a classical method of thresholding and is very easy to implement as well as it is very effective at gray level scale which is why it is mostly used in image segmentation.

It only selects the threshold value to reduce the intra-class threshold values of only black and white pixels.

For segmentation of MRI images, Otsu method of thresholding is applied in this chapter. In this method the variance of the object and its background's both are exclusive of the complete image variance. The modified and improved Otsu method offers superior outcomes and proficient thresholding for both unimodal distribution and bimodal distribution of image with histogram. Otsu's method is centered on the discriminant analysis and it principally distributes the image into two independent classes 'B_1' and 'B_2'$_2$ at gray levels 'm' such that $X_{11} = \{0, 1, 2, 3, \ldots, r\}$ and $X_2 = \{r + 1, r + 2, \ldots, s\text{-}1\}$ here, 'n' is the overall value of gray levels of the MR image. Assume total pixels number in the given MR image is 'k' hence the ith pixel of the image is represented by k_i.

The gray level occurrence probability is defined by

$$P = \frac{k_i}{k} \tag{9}$$

'B_1' and 'B_2' are two classes of MR image demonstrating the region of interest of tumor detection and the background of the image, hence the probabilities of the classes X1 and X2 are given as:

$$\begin{aligned} P_1(r) &= \sum_{i=0}^{r} P_i \\ P_2(r) &= \sum_{i=r+1}^{s-1} P_i = 1 - P_1(r) \end{aligned} \tag{10}$$

These two classes B_1 and B_2 mean intensity values are given as:

The mean gray level of the Image and its background pixels are given by

$$m_1(r) = \sum_{i=0}^{r} i \cdot P\left(\frac{i}{X_1}\right) = \sum_{i=0}^{r} i \cdot \frac{P(i)}{P(X_1)} \tag{11}$$

$$m_1(r) = \frac{1}{P_1(r)} \sum_{i=0}^{r} i \cdot P_i \tag{12}$$

Similarly

$$m_2(r) = \sum_{i=r+1}^{s-1} i \cdot P\left(\frac{i}{X_2}\right)$$

$$m_2(r) = \frac{1}{P_2(r)} \sum_{i=r=1}^{s-1} i \cdot P_i \tag{13}$$

The variance of the Image and its background pixels are given by

$$\sigma_1^2(r) = \frac{\sum_{i=0}^{r} (i - m_1(r))^2 \cdot P_i}{\sum_{i=0}^{r} P_i} \tag{14}$$

$$\sigma_1^2(r) = \frac{1}{P_1(r)} \cdot \sum_{i=0}^{r} (i - m_1(r))^2 \cdot P_i \tag{15}$$

If σ_T^2 denotes the total variance and σ_B^2 represents between class variance, by maximizing between class variance an optimal threshold T* can be obtained.

$$T^* = Arg\left(\begin{array}{c} Max \\ 0 < r < s - 1\left(\frac{\sigma_B^2}{\sigma_T^2}\right) \end{array}\right) \tag{16}$$

This optimal threshold obtained from Eq. (16) is used for converting the images into binary form.

5 RBFNN Classifier

The RBFNN (Radial Basis Function Neural Network) classifier is projected for the classification of the brain tumors in three classes from the segmented images. The learning rate of the RBFNN classifier is greater compared to other neural classifier, that's why we have selected RBFNN as classifier. It offers numerous advantages over other classifiers such as robust to acceptance level of input noise, simple implementation, fast online learning rate and robust to excellent generalization [49].

RBFNN contains multilayered neurons with many hidden layers and also it is a hybrid non-linear neural network.

The uppermost layer of RBFNN is called the input layer which receives the input features and transfers it to the hidden layer, then the activation function is formed grounded on the radial basis function connected in the second stage (hidden layers).

Depending on the nature of the classification, the output is produced by the RBFNN classifier at the output (O/P) layer, which is a linear combination of the hidden units and activation function derived from the second stage [24].

Assuming weight vector $W = [w_1, w_2, w_3, \ldots w_n]^T$ and the input vector is represented by $h(x)$ and it is the function of "x" with $\emptyset = 1$. The RBFNN performance

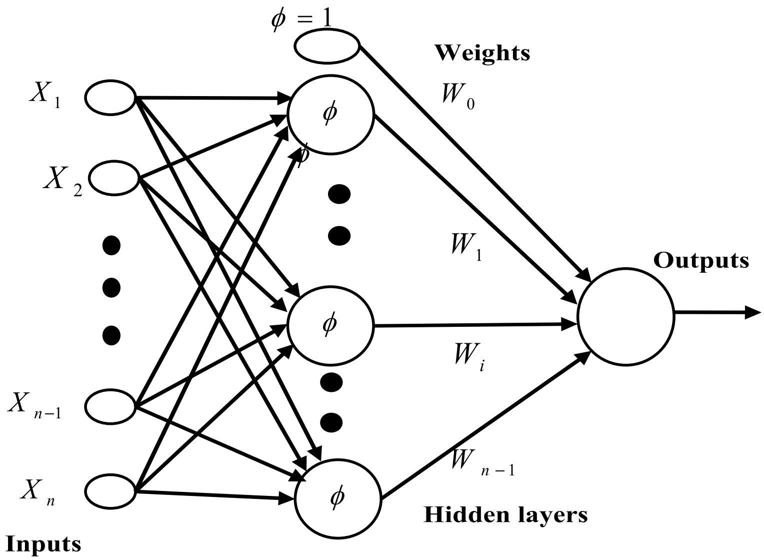


Fig. 4 The basic structure of proposed classifier RBFNN

behavior is expressed by the function

$$\check{f}(x, w) = \sum_{i=1}^{n1} w_i \varphi_i(x) \tag{17}$$

where, $\varphi_i = G(x - \mu_i), i = 1, 2, \ldots n_1$.

Generally in image classification problems, Gaussian function is used as the basis function in the RBFNN [15]. The RBFNN utilized in this chapter to categorize the O/P into 3 classes of Brain MRI normal, benign and malignant class [38]. The basic structural arrangement of RBFNN classifier is shown in the Fig. 4.

5.1 Performance Evaluation Assessment Metrics (PEAM)

To evaluate and verify the performance of the classifiers, many evalution metrics are used. Depending upon the application and types of classifier used the PEAEM are computed and selected. For this work, we have carefully chosen seven types of PEAM which were calculated from the confusion matrix of the classifier. The classification result is calculated from the confusion matrix and the PEAM values were computed. The utilized PEAM are Recall (Rec), Precision (Pr), F1-score (F1), Specificity (Sp) and Classification error rate (CER) accuracy (Acc) and overall accuracy (OAcc). The mathematical expression to calculate the PEAM are given as [44]:

$$Precision\,(\mathrm{Pr}) = \frac{TP}{TP + FP}$$

$$Recall\,(Re) = \frac{TP}{TP + FN}$$

$$Specificity\,(Sp) = \frac{TN}{TN + FP}$$

$$F - Score = \frac{2 * Precision * Recall}{Precesion + Recall}$$

$$Accuracy\,(Acc) = \frac{TP + TN}{TP + FN + TN + FP}$$

where, TP = True Positive, FP = False Positive, TN = True Negative and FN = False Negative.

6 Experimental Result and Discussion

The experiment was setup and performed on MATLAB environment in software package 2015a.

The 1st step was to preprocess the BMRI images before segmenting the tumor from it. In preprocessing stage, the BMRI images were denoised using hybrid DWT-ICA method which is explained in detail in the Sect. 3.2. After denoising of MR images, in the 2nd step of preprocessing, the BMRI images are resized, since these are not homogeneous in nature. The shape of resized images are of 256 × 256 × 3, where 256 is the height and width of the image and 3 is the number of channel for color images (RGB).

6.1 BMRI Image Segmentation Process

After preprocessing of BMRI images the tumor segmentation task is performed using Otsu thresholding method. The steps involved for segmenting the tumor mask from BMRI images is explained in this section. Initially the preprocessed raw BMRI images are loaded into the MTLAB which are denoised and resized. The Otsu thresholding method is employed to obtain the optimized threshold and binary images are obtained, by applying steps 2–10. The threshold (binary) images are then converted into binary inverted images that means the inverse of binary where the black and white is inverted into white and black respectively, from step 11. Then applied AND operation on the binary inverted images to find the canny edges and the tumor mask area, from step 12. And at final step 13, we have found the contour of the masked

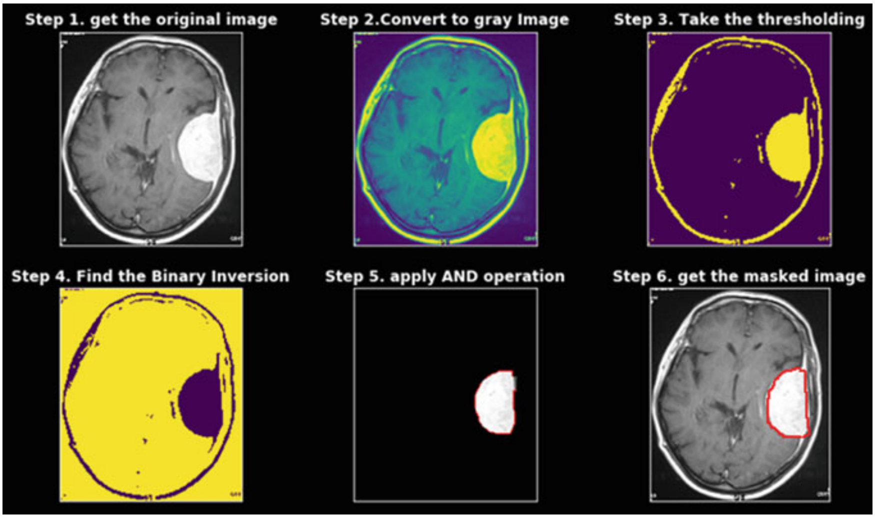


Fig. 5 Segmentation steps to find tumor mask from BMRI images

tumor and merged with the original images with transparent background to show brain tumor mask with contour. The steps involved in the tumor segmentation of BMRI images is demonstrated in Fig. 5 and detailed as:

1. Calculate the histogram of the preprocessed images obtained by applying the DWT-ICA denoising techniques (Sect. 3.2), from Eq. (8).
2. All the calculated histogram values are added together.
3. Initialize max = 0, the probability of every intensity level are calculated and applied to all further values from 2–255.
4. Each class probability of the MR images are computed using Eq. (10).
5. Mean values for both the classes are calculated from Eqs. (11) to (13).
6. The variance or sigma (σ) value between the classes is obtained using Eqs. (14) and (15).
7. The (σ > max) condition is checked, if false, repeat from the step 3 to 7.
8. If true, update the max value with σ, (max = σ).
9. Preferred value of threshold (T*) is obtained corresponding to maximum σ value using Eq. (16).
10. Applying the desired T* values the images are converted into binary form.
11. From the binary images, its inversion is obtained.
12. Then Canny edges are obtained and AND operation is applied.
13. The contours of the images are determined and segmented tumor is obtained.

6.2 Feature Extraction

After the segmentation of tumor from BMRI images using Otsu thresholding method, we extracted the features using DWT. Feature extraction is very much important in machine learning and image processing task where feature stands for the most useful information extracted from the data to feed as the input to the classifier. Good features always recommended to extracted to enhance the efficiency of the classifier.

The feature extraction of Brain MRI (BMRI) images was accomplished by means of Discrete WT. We have extracted two types of features from the segmented and preprocessed images that are statistical features, and morphological features. The images can be differentiated based on their pixel values features, statistical values of pixels also shows impactful result in classification. The extracted statistical features are Mean, Median, Variance, PSD, SD, RMS, and Correlation and morphological features are entropy, energy, contrast, smoothness, skewness, homogeneity. Hence we extracted total thirteen types of features based on morphology of the image and the statics. Out of 13 features extracted from the BMRI dataset, 5 are morphology based features and 8 are statics based features.

6.3 Feature Normalization

Feature Normalization (FN) is a process of scaling the data of a fixed range generally between 0 and 1. Since we have used various types of features based on morphology of the images and also statistics based feature hence it is needed to scale up the data between ranges 0–1. The reason for using FN for this work to converge the gradient descent faster as compared to without normalization. We have used min-max normalization for our work, which is given by

$$\bar{y} = \frac{y - y_{\min}}{y_{\max} - y_{\min}}$$

where 'y' is the feature vector, y_{min} and y_{max} is the minimum and maximum values of feature vector respectively.

6.4 Data Distribution

To perform the experiment on proposed classifier, FFNN and BPNN the BMRI dataset were used. Distribution of dataset is also important to train, validate and test the classifier properly. If the data is not properly distributed, for example if we trained the network on less data it won't be able to learn and if provide more data for training then there may be the problem of overfitting.

The total 433 BMRI images are available in the dataset including 105 benign, 105 malignant and 223 normal without any tumor. The entire BMRI dataset was distributed into three parts training, testing and validation set. We have used 10% images (43) for validation, 25% of remaining ($433 - 43 = 390$) i.e. 98 images for testing and remaining ($433 - 43 - 98$) 292 images for training the neural classifiers. The 292 images were used to train the classifiers and to verify it we have used 43 images (validation). The test dataset consisting 98 images, which is not exposed in the training process, were used for testing the models.

6.5 Classification Result

The final experimentation was performed using proposed classifier RBFNN to classify the BMRI images into three categories normal, benign and malignant, on the segmented images. The complete flow chart of employed segmentation and classification method for tumor detection from brain MR images are illustrated in Fig. 6. The extracted features (statistical and morphological) by applying DWT is used as the input to the RBFNN classifier. The proposed classifier result was compared with other two classifiers FNNN and BPNN on the same dataset. 1st classification experiment was performed on the RBFNN proposed classifier on training (292) and validation (43) dataset. The result was verified and tested on the test dataset (98), for unbiased performance verification, this dataset was not exposed during training.

The test data (98) contains 46 normal BMRI images, 26 benign tumor images and 26 malignant tumor images and also the confusion matrix on the test dataset was tabulated for three class. The 2nd experiment was performed on the same dataset using BPNN and 3rd using FFNN. The testing, verification and predication was experimented on the test dataset of 98 BMRI images. Table 1 arranges the confusion matrix for all class normal (N), benign (B) and Malignant (M) using all three classifiers along with TP, FN, FP, TN, actual class and total test images. The bold test represents the true positive (TP) class for each class and classifiers.

Table 2 visualizes the all seven Performance Evaluation Assessment Metrics (PEAM) for propose RBFNN, BPNN and FFNN classifiers. The PEAM values in percentage for each class were computed using classification metrics shown in Table 2. The PEAM metrics includes percent precision (%Pr), percent recall (%Rec), percent specificity (%Sp), percent F1-score (%F1), percent accuracy (% Acc), classification error rate (% CER), and percent overall accuracy (% OAcc).

The proposed classifier performance was compared with BPNN and FFNN using PEAM in percent for each class and also the average of PEAM is calculated for each classifiers. The result of the experiment suggests that the proposed classifier (RBFNN) on the test data provides 100% OAcc compared to 97.96% OAcc for the BPNN classifier and 95.92% OAcc for the FNNN classifier, shown in Fig. 7.

The average % accuracy for all three class normal, benign and malignant using RBFNN, BPNN and FFNN is shown in Fig. 8. From figure, it is also visualized

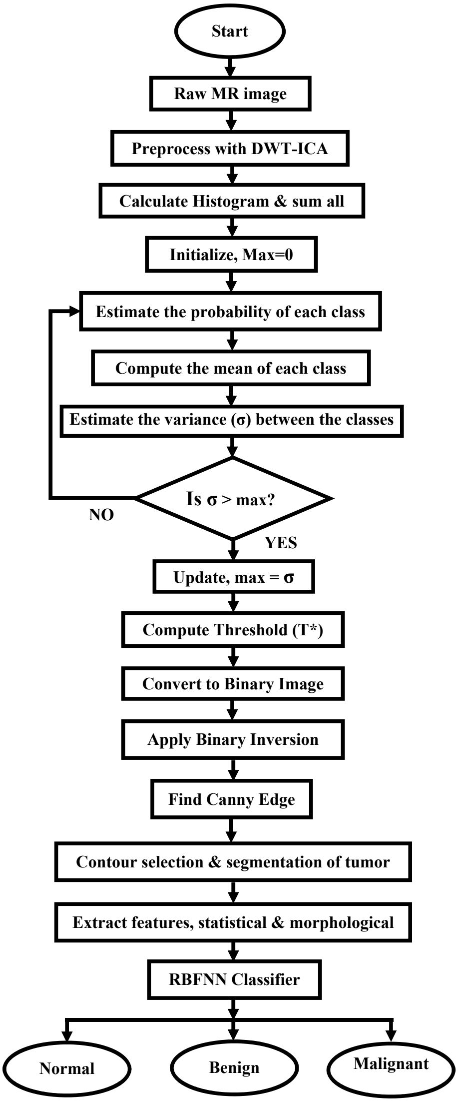


Fig. 6 Flowchart for brain tumor segmentation and classification using proposed methodology

that the average accuracy of RBFNN classifier is 100% as compared to 98.64% and 97.28% for BPNN and FFNN classifier respectively.

The % average precision for all three classes using FFNN, BPNN and RBFNN classifiers are shown in Fig. 9. The Average precision achieved using proposed classifier is 100%, 97.62% for BPNN and 95.24% for FFNN respectively.

The % average CER for all class using all three classifiers FFNN, BPNN and RBFNN are shown in Fig. 10. The average CER of RBFNN classifier is 0%, 1.36% and 2.72% using BPNN and FFNN respectively.

Table 1 Confusion matrix of all categories and all three classifiers

Classifiers	Types of class	N	B	M	TP	FN	FP	TN	Actual Class	Total
FFNN	Normal	**42**	0	0	42	4	0	52	46	98
	Benign	2	**26**	0	26	0	2	70	26	98
	Malignant	2	0	**26**	26	0	2	70	26	98
BPNN	Normal	**44**	0	0	44	2	0	52	46	98
	Benign	2	**26**	0	26	0	2	70	26	98
	Malignant	0	0	**26**	26	0	0	72	26	98
RBFNN	Normal	**46**	0	0	46	0	0	52	46	98
	Benign	0	**26**	0	26	0	0	72	26	98
	Malignant	0	0	**26**	26	0	0	72	26	98

Table 2 PEAM results for FFNN, BPNN and proposed (RBFNN) classifiers

Classifiers	Type of Class	% Pr	% Rec	% Sp	%F1	Acc (%)	% CER	% OAcc
FFNN	Normal	100.0	91.3	100.0	95.5	95.9	4.1	**95.92**
	Benign	92.9	100.0	97.2	96.3	98.0	2.0	
	Malignant	92.9	100.0	97.2	96.3	98.0	2.0	
Average		**95.2**	**97.1**	**98.1**	**96.0**	**97.3**	**2.7**	
BPNN	Normal	100.0	95.7	100.0	97.8	98.0	2.0	**97.96**
	Benign	92.9	100.0	97.2	96.3	98.0	2.0	
	Malignant	100.0	100.0	100.0	100.0	100.0	0.0	
Average		**97.6**	**98.6**	**99.1**	**98.0**	**98.6**	**1.4**	
RBFNN	Normal	100.0	100.0	100.0	100.0	100.0	0.0	**100**
	Benign	100.0	100.0	100.0	100.0	100.0	0.0	
	Malignant	100.0	100.0	100.0	100.0	100.0	0.0	
Average		**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**0.0**	

The % average Recall using all three classifiers is plotted by bar plot and shown in the Fig. 11. Figure also show that proposed classifier provides 100% average % Recall and 98.55%, 97.10% using BPNN and FFNN respectively.

The average F1-Score of all three classes normal (N), benign (B) and malignant (M) in percentage using FFNN, BPNN and RBFNN classifiers are shown in Fig. 12. The Average F1-score on test dataset using proposed, BPNN and FFNN are 100%, 98.02% and 96.02% respectively.

The average specificity of all three classes normal (N), benign (B) and malignant (M) in percentage using FFNN, BPNN and RBFNN classifiers are shown in Fig. 13.

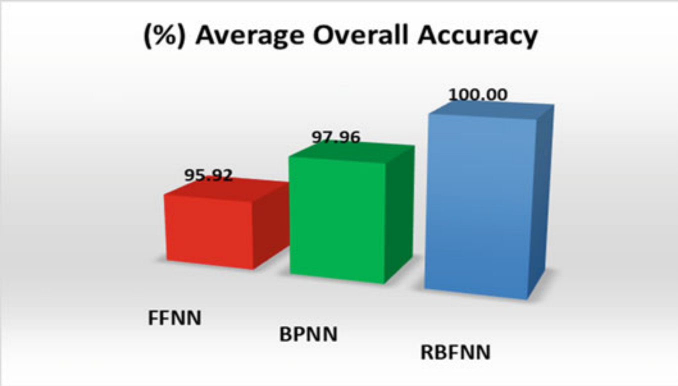


Fig. 7 Average percentage of overall accuracy using FFNN, BPNN and RBFNN classifiers

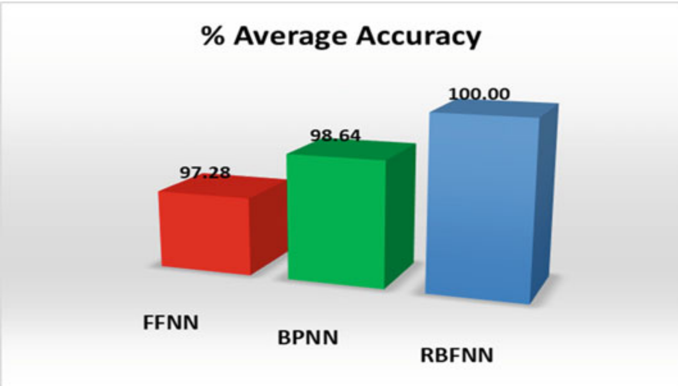


Fig. 8 Average percentage of accuracy using FFNN, BPNN and RBFNN classifiers

The Average Specificity percentage on test dataset using proposed, BPNN and FFNN are 100%, 99.07% and 98.15% respectively.

After comparing the performance result of each classifier based on PEAM it was observed that the proposed classifier on test data provides 100% result in all aspects. Also all evaluation metrics %Pr, % Rec, % Sp, %F1, %Acc, % CER, and %OAcc are 100% using proposed classifier.

6.6 *Result Analysis*

Results have been analyzed and compared with state-of-the-art methods based on the classification performance. Table 3 shows the analysis, research findings, and

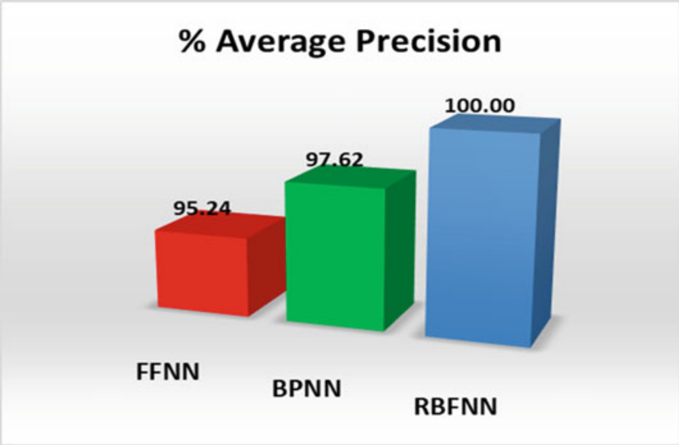


Fig. 9 Average percentage of precision using FFNN, BPNN and RBFNN classifiers

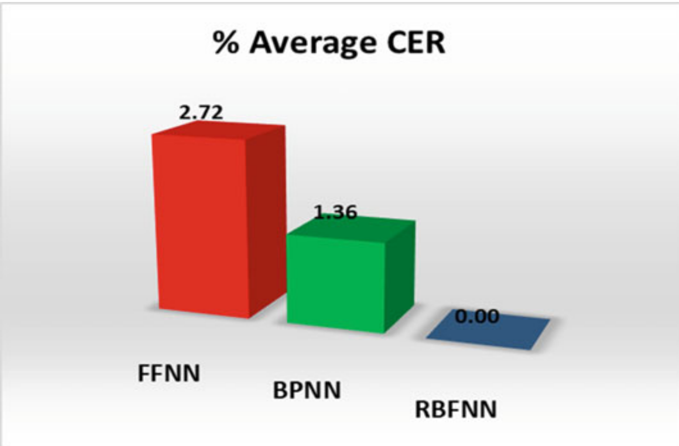


Fig. 10 Average percentage of CER using FFNN, BPNN and RBFNN classifiers

research gap comparisons by some present and past researchers with the proposed method based on the database Size, performance assessments such as accuracy (Acc), sensitivity (Se), and specificity (Sp) of classification. It was observed that the proposed model using Otsu + canny + RBFNN based classification performed best as compared to all the state-of-art method. The image dataset used for validation of the proposed method is also satisfactory with the use of 443 MR images, as well as the performance (%) achieved in terms of accuracy (Acc), sensitivity (Se) and specificity (Sp) is 100%. Hence, after analyzing the segmentation and classification result of the proposed method it was found the RBFNN classifier, with input as statistical

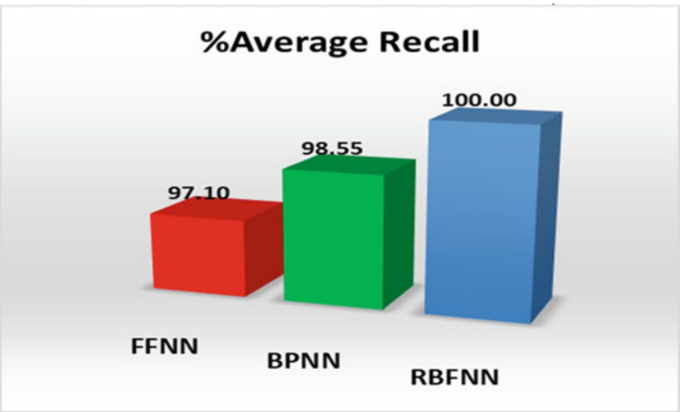


Fig. 11 Average percentage of recall using FFNN, BPNN and RBFNN classifiers

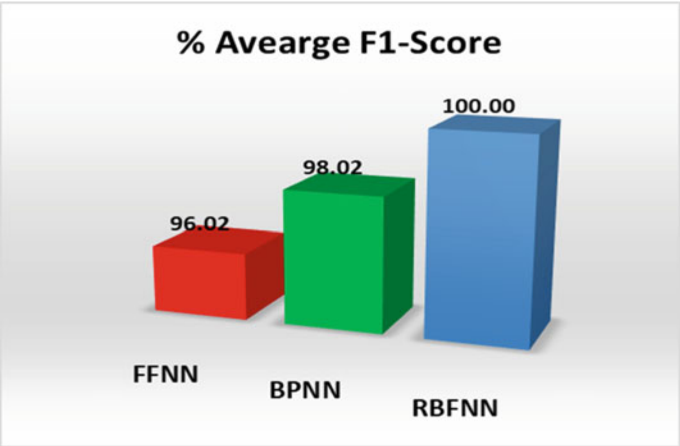


Fig. 12 Average percentage of F1-score using FFNN, BPNN and RBFNN classifiers

and morphological features from the Otsu and canny edge segmented MR images, is very efficient and accurate. Clarification

7 Conclusion and Future Scope

The automatic segmentation and classification of tumor from BMRI images using RBFNN classifier has been presented in this chapter. The proposed method of segmentation and classification was divided into five main parts, Preprocessing, Segmentation, post processing, feature extraction and classification.

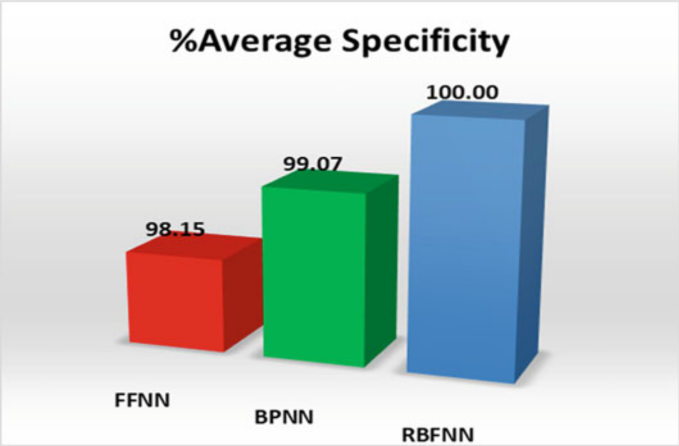


Fig. 13 Average percentage of specificity using FFNN, BPNN and RBFNN classifiers

In preprocessing stage, denoising of BMRI is done using hybrid DWT-ICA method and then denoised images are resized of shape 256 × 256 × 3. The segmentation of brain tumor mask is performed using Otsu thresholding method. The feature from the tumor segmented images were extracted using DWT. Total 13 types of feature were extracted of two types statistical (8) and morphological (5) using DWT method. In post-processing stage, the feature normalization were performed on feature vector using min-max approach. RBFNN is used for the classification of brain tumors and the evaluation of the performance assessment was done using 7 types of PEAM. The proposed classifier (RBFNN) result was also compared with BPNNC and FFNN based on performance evaluation metrics. Experimentation was performed on 433 BMRI dataset which consists three types of images 233 normal, 105 benign and 105 malignant. From the experimental result it was found that proposed classifier RBFNN outperform FFNN and BPNN in all aspects of assessment metrics and shows 10% over all accuracy on test dataset. The proposed classifier on the segmented features is very efficient and accurate for the classification of tumor from BMRI images.

In the future scope of work, intensive learning based hybrid deep neural network (DNN) models are proposed on large datasets with great accuracy. Computation or simulation time also plays an important role for evaluating classification performance measure, particularly in the case of medical data, so the fast DNN model for brain tumor detection is proposed for future work. Also, both 2D and 3D types of MRI datasets will be used not only to detect tumors, but also to identify specific tumor types in future work of this chapter.

Table 3 Comparative analysis of proposed technique with state-of-art methods

Literature	SM	CM	DS	Performance (%)		
				Acc	Se	Sp
Nachimuthu and Baladhandapani [30]	DFM, MRS	ELM-IPSO	35 (C)	99.2	98	95
Shanthakumar and Ganeshkumar [42]	RGM	ANFIS	200 (I)	87	81	–
Vishnuvarthanan et al. [48]	SOM + FKM	–	4 (P) + HBR	96.2	87.2	
Ilhan and Ilhan [16]	NTM	–	100 (I)	96	94.3	100
Al-Dmour and Al-Ani [4]	CFT	KMC, FCM, SOM	28 (P)	98.1	79.7	99
Deepa and Sam Emmanuel [10]	–	AFBNN	81 (I)	99.8	97.2	99.9
Salem Ghahfarrokhi and Khodadadi [40]	FCMC	Pattern Net	233 (P)	98.9	96.5	96.7
Virupakshappa and Amarapur [47]	CMLS	AANN	274 (I)	98	96	98
H. M. Rai and K. Chatterjee	**Otsu + Canny**	**RBFNN**	**443 (I)**	**100**	**100**	**100**

SM - Segmentation Methods, *CM* - Classification Methods, *DS* - Dataset Size, *DFM* - Data fusion of MRI, *MRS* - magnetic resonance spectroscopy, *ELM-IPSO* - extreme learning machineimproved particle swarm optimization, *RGM* - Region growing method, *ANFIS* - adaptive neuro fuzzy inference system, *SOM* - self-organizing map (SOM), *FKM* - fuzzy K means, *HBR* - Harvard brain repository, *NTM* - New thresholding method, *CFT* - clustering fusion technique, *KMC* - k-means clustering, *FCM* - fuzzy c-means, *AFBNN*- adaptive firefly backpropagation neural network, *FCMC* - fuzzy c-means clustering, *MLS* - modified Level set, *AANN* - Adaptive Artificial Neural Network, *C* - Cases, *I* - Images, *P* - Patients

References

1. ADNI: ADNI Alzheimer's Disease Neuroimaging Initiative (2003). http://adni.loni.usc.edu/
2. Ahmed, H.M., Youssef, B.A.B., Elkorany, A.S., Saleeb, A.A., Abd El-Samie, F.: Hybrid gray wolf optimizer–artificial neural network classification approach for magnetic resonance brain images. Appl. Opt. **57**(7), B25 (2018). https://doi.org/10.1364/ao.57.000b25
3. Akansu, A.N., Haddad, R.A.: Chapter 6 Wavelet transform. In: Multiresolution Signal Decomposition, pp. 391–442. Academic Press (2001). https://doi.org/10.1016/B978-0-12-047141-6.50006-9
4. Al-Dmour, H., Al-Ani, A.: A clustering fusion technique for MR brain tissue segmentation. Neurocomputing **275**, 546–559 (2018). https://doi.org/10.1016/j.neucom.2017.08.051
5. Alfonse, M., Salem, A.M.: An automatic classification of brain tumors through MRI using support vector machine. Egypt. Comput. Sci. J. **40**(03), 1110–2586 (2016)

6. Anjali, R., Priya, S.: An efficient classifier for brain tumor classification. Int. J. Comput. Sci. Mobile Comput. **6**(8), 40–48. www.ijcsmc.com (2017)
7. Balafar, M.A., Ramli, A.R., Saripan, M.I., Mashohor, S.: Review of brain MRI image segmentation methods. Artif. Intell. Rev. **33**(3), 261–274 (2010). https://doi.org/10.1007/s10462-010-9155-0
8. CDRC: MRI database "Charak Diagnostic & Research Center." http://charakdnrc.com/mri.htmlb (2012)
9. Constantinides, C.: Magnetic Resonance Imaging: The Basics. CRC Press (Taylor & Francis) (2016)
10. Deepa, A.R., Sam Emmanuel, W.R.: An efficient detection of brain tumor using fused feature adaptive firefly backpropagation neural network. Multimedia Tools Appl. **78**(9), 11799–11814 (2019). https://doi.org/10.1007/s11042-018-6731-9
11. Devkota, B., Alsadoon, A., Prasad, P.W.C., Singh, A.K., Elchouemi, A.: Image segmentation for early stage brain tumor detection using mathematical morphological reconstruction. Procedia Comput. Sci. **125**(2009), 115–123 (2018). https://doi.org/10.1016/j.procs.2017.12.017
12. Fayed, L.: Differences between a malignant and benign tumor. Verywell 1–5 (2020). https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240
13. Frellick, M., Vega, C.P.: How big a problem is misdiagnosis in medicine? Medscape. https://www.medscape.org/viewarticle/933116 (2020, February 7)
14. Gupta, N., Bhatele, P., Khanna, P.: Glioma detection on brain MRIs using texture and morphological features with ensemble learning. Biomed. Signal Process. Control **47**, 115–125 (2019). https://doi.org/10.1016/j.bspc.2018.06.003
15. Hajek, M.: Neural Networks, pp. 10–13 (2005). https://doi.org/10.1016/j.neunet.2004.10.001
16. Ilhan, U., Ilhan, A.: Brain tumor segmentation based on a new threshold approach. Procedia Comput. Sci. **120**, 580–587 (2017). https://doi.org/10.1016/j.procs.2017.11.282
17. Inje, B., Kumar, S., Nayyar, A.: Swarm intelligence and evolutionary algorithms in healthcare and drug development. In: Swarm Intelligence and Evolutionary Algorithms in Healthcare and Drug Development, pp. 1–18. Chapman and Hall/CRC (2019). https://doi.org/10.1201/9780429289675
18. Jaya, J., Thanushkodi, K.: Exploration on selection of medical images employing New Transformation Techniques. IJCSI Int. J. Comput. Sci. **4**(3) (2010)
19. Jeevakala, S., Brintha Therese, A.: Sharpening enhancement technique for MR images to enhance the segmentation. Biomed. Signal Process. Control **41**, 21–30 (2018). https://doi.org/10.1016/j.bspc.2017.11.007
20. Ji, Z., Huang, Y., Sun, Q., Cao, G.: A spatially constrained generative asymmetric Gaussian mixture model for image segmentation. J. Vis. Commun. Image Represent. **40**, 611–626 (2016). https://doi.org/10.1016/j.jvcir.2016.08.001
21. Joshi, D., Choubey, N., Kumari, R.: Brain tumour diagnosis. In: Swarm Intelligence and Evolutionary Algorithms in Healthcare and Drug Development, pp. 45–64. Chapman and Hall/CRC (2019)
22. Kalaiselvi, T., Kumarashankar, P., Sriramakrishnan, P., Karthigaiselvi, S.: Brain tumor detection from multimodal MRI brain images using pseudo coloring processes. Procedia Comput. Sci. **165**, 173–181 (2019). https://doi.org/10.1016/j.procs.2020.01.094
23. Kanmani, P., Marikkannu, P.: MRI brain images classification: a multi-level threshold based region optimization technique. J. Med. Syst. **42**(4), 1–12 (2018). https://doi.org/10.1007/s10916-018-0915-8
24. Korürek, M., Doğan, B.: ECG beat classification using particle swarm optimization and radial basis function neural network. Expert Syst. Appl. **37**(12), 7563–7569 (2010). https://doi.org/10.1016/j.eswa.2010.04.087
25. Kumar, A., Krishnamurthi, R., Nayyar, A., Sharma, K., Grover, V., Hossain, E.: A novel smart healthcare design, simulation, and implementation using healthcare 4.0 processes. IEEE Access **8**, 118433–118471 (2020). https://doi.org/10.1109/ACCESS.2020.3004790

26. LaMontagne, P., Benzinger, T., Morris, J., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., Raichle, M., Cruchaga, C., Marcus, D.: OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv (2019). https://doi.org/10.1101/2019.12.13.19014902
27. Mahapatra, B., Krishnamurthi, R., Nayyar, A.: Healthcare models and algorithms for privacy and security in healthcare records. In: Security and Privacy of Electronic Healthcare Records: Concepts, Paradigms and Solutions, pp. 183–221. Institution of Engineering and Technology (2019). https://doi.org/10.1049/PBHE020E_ch8
28. Mahapatra, B., Nayyar, A.: Swarm intelligence and evolutionary algorithms for cancer diagnosis. In: Swarm Intelligence and Evolutionary Algorithms in Healthcare and Drug Development, pp. 19–44. Chapman and Hall/CRC (2019)
29. Mirajkar, G., Barbadekar, B.: Automatic segmentation of brain tumour from MR images using undecimated wavelet transform and Gabor wavelets. IEEE Int. Conf. Electron. Circuits Syst. 702–705 (2010)
30. Nachimuthu, D.S., Baladhandapani, A.: Multidimensional texture characterization: on analysis for brain tumor tissues using MRS and MRI. J. Digit. Imaging **27**(4), 496–506 (2014). https://doi.org/10.1007/s10278-013-9669-5
31. Nagalkar, V.J., Asole, S.S.: Brain tumor detection using digital image processing based on soft computing. Image Process. **3**(3), 102–105 (2012)
32. Nayyar, A., Puri, V., Nguyen, N.G.: BioSenHealth 1.0: a novel internet of medical things (IoMT)-based patient health monitoring system. In: International Conference on Innovative Computing and Communications, vol. 55, pp. 155–164. Springer Singapore (2019). https://doi.org/10.1007/978-981-13-2324-9_16
33. Nazir, M., Wahid, F., Ali Khan, S.: A simple and intelligent approach for brain MRI classification. J. Intell. Fuzzy Syst. **28**(3), 1127–1135 (2015). https://doi.org/10.3233/IFS-141396
34. NBTS.: Quick brain tumor facts. National Brain Tumor Society. https://braintumor.org/brain-tumor-information/brain-tumor-facts/ (2016)
35. OASIS.: OASIS brains—open access series of imaging studies. http://www.oasis-brains.org/ (2019)
36. Pramanik, P.K.D., Solanki, A., Debnath, A., Nayyar, A., El-Sappagh, S., Kwak, K.S.: Advancing modern healthcare with nanotechnology, nanobiosensors, and internet of nano things: taxonomies, applications, architecture, and challenges. IEEE Access **8**, 65230–65266 (2020). https://doi.org/10.1109/ACCESS.2020.2984269
37. Rai, H.M., Chatterjee, K.: Hybrid adaptive algorithm based on wavelet transform and independent component analysis for denoising of MRI images. Measure. J. Int. Measure. Confederat. **144**, 72–82 (2019). https://doi.org/10.1016/j.measurement.2019.05.028
38. Rai, H.M., Trivedi, A., Chatterjee, K., Shukla, S.: R-peak detection using Daubechies wavelet and ECG signal classification using radial basis function neural network. J. Instit. Eng. (India): Ser. B **95**(1), 63–71 (2014). https://doi.org/10.1007/s40031-014-0073-4
39. Saba, L.: Magnetic Resonance Imaging Handbook (2016)
40. Salem Ghahfarrokhi, S., Khodadadi, H.: Human brain tumor diagnosis using the combination of the complexity measures and texture features through magnetic resonance image. Biomed. Signal Process. Control **61**, (2020). https://doi.org/10.1016/j.bspc.2020.102025
41. Selvaraj, D., Dhanasekaran, R.: A review on tissue segmentation and feature extraction of MRI brain images **4**(10), 1313–1332 (2013)
42. Shanthakumar, P., Ganeshkumar, P.: Performance analysis of classifier for brain tumor detection and diagnosis. Comput. Electr. Eng. **45**, 302–311 (2015). https://doi.org/10.1016/j.compeleceng.2015.05.011
43. Thapaliya, K., Pyun, J.Y., Park, C.S., Kwon, G.R.: Level set method with automatic selective local statistics for brain tumor segmentation in MR images. Comput. Med. Imaging Graph. **37**(7–8), 522–537 (2013). https://doi.org/10.1016/j.compmedimag.2013.05.003
44. Tharwat, A.: Classification assessment methods. Appl. Comput. Inf. (2018). https://doi.org/10.1016/j.aci.2018.08.003

45. The Indian Express: World Brain Tumour Day 2019: know the symptoms, risk factors and treatment. https://indianexpress.com/article/lifestyle/health/world-brain-tumour-day-2019-symptoms-risk-factors-treatment-5770587 (2019)
46. Varuna Shree, N., Kumar, T.N.R.: Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network. Brain Inf. **5**(1), 23–30 (2018). https://doi.org/10.1007/s40708-017-0075-5
47. Virupakshappa, Amarapur, B.: Computer-aided diagnosis applied to MRI images of brain tumor using cognition based modified level set and optimized ANN classifier. Multimed. Tools Appl. **79**(5–6), 3571–3599 (2020). https://doi.org/10.1007/s11042-018-6176-1
48. Vishnuvarthanan, G., Rajasekaran, M.P., Subbaraj, P., Vishnuvarthanan, A.: An unsupervised learning method with a clustering approach for tumor identification and tissue segmentation in magnetic resonance brain images. Appl. Soft Comput. J. **38**, 190–212 (2016). https://doi.org/10.1016/j.asoc.2015.09.016
49. Yu, H., Xie, T.: Advantages of radial basis function networks for dynamic system design. IEEE Trans. … **58**(12), 5438–5450 (2011). https://doi.org/10.1109/TIE.2011.2164773
50. Zeinali, Y., Story, B.A.: Competitive probabilistic neural network. Integrat. Comput.-Aided Eng. **24**(2), 105–118 (2017). https://doi.org/10.3233/ICA-170540

Automatic Localization of Optic Disc in Retinal Fundus Image Based on Unsupervised Learning

J. Prakash and B. Vinoth Kumar

Abstract Automated localization of the optic disc is an essential phase in automated retinal imaging system, which are used in determining the effect of diabetic retinopathy and glaucoma at their initial phase. In this chapter an unsupervised learning technique called clustering methods are used in localization of optic disc. The clustering methods like DBSCAN clustering, Hierarchical Clustering, Fuzzy C Mean and K-Mean clustering are used in optic disc localization. The pre-processing techniques like resizing to avoid images high computation requirement, selection of green color band for processing and smoothing using morphological and gaussian blur operator are performed. Then the optic disc is localized by applying clustering methods by filtering the maximum intensity cluster. The performance of clustering methods was evaluated on three publicly available retinal fundus image databases DRIONSDB dataset containing 110 fundus images, DIARETDB1dataset containing 89 fundus images and High-Resolution Fundus dataset containing 45 fundus images for 10 and 20 iterations respectively. Positive Predicted Value, False Discovery Rate, Accuracy, Error rate, F1 Score, Precision and Recall was determined to analyze the performance of clustering methods. The results infer that the performance of clustering techniques on High Resolution Fundus dataset is better than the other datasets.

1 Introduction

Worlds human population is increasing rapidly, with the increasing population the many people are getting affected by eye disease. Many researchers have identified that the major source of the blindness amongst the world population is because of glaucoma. It affects the optic nerve in the human eye by causing loss of vision. It is

J. Prakash (✉)
Department of CSE, PSG College of Technology, Coimbatore, India
e-mail: jpk.cse@psgtech.ac.in

B. Vinoth Kumar
Department of IT, PSG College of Technology, Coimbatore, India
e-mail: bvk.it@psgtech.ac.in

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_11

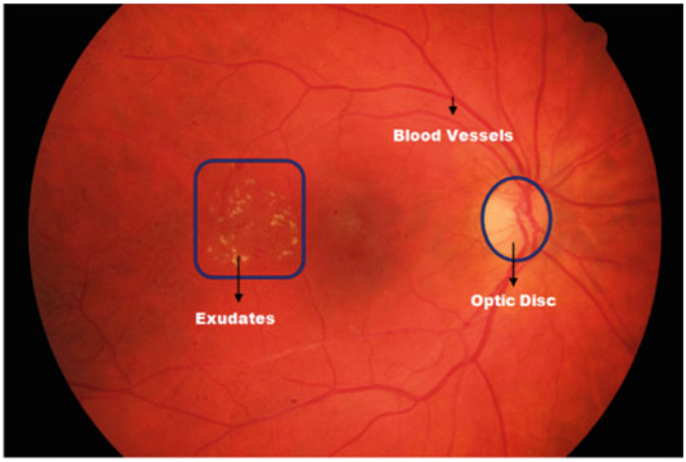


Fig. 1 Structure of fundus image

identified that glaucoma will be severe for the people aged above 40 years and will be very severe for the people aged above 60 years. If the glaucoma is identified at the later stages curing it will become a difficult task. So, the glaucoma is needed to be identified at the earlier stages to control the vision loss. The people will be seen with no symptoms until the human vision get affected which make the detection of eye disease more challenging. The glaucoma can be diagnosed by determining the damage in optic nerve head or by assessing the visual field. Optic disc is a part in the human eye, from where all the optic nerves connect to. It is vertically oval and the dimensions ranging from the average of 1.92 mm vertical dimension and 1.76 mm horizontal dimension [1]. Figure 1 represent the structure of fundus eye image.

The optic disc detection is a vital part of glaucoma diagnosis. In detection of optic disc, the portion of it need to be segmented for glaucoma diagnosis. Optic disc will be varying in sizes and looks bright yellowish color with the blood vessels around it. Cup disc ratio are crucial for diagnosis of disease. The optic disc shall have normal cupping or sometimes they may be flat. Determining the cup disc ratio is an essential part in glaucoma diagnosis. Glaucoma or any other pathology is confirmed if the cup disc ratio is high. If a person is normal and not affected by glaucoma then his vertical cup disc ratio will be lesser than horizontal cup disc ratio. Determining the cup disc ratio is a tedious and laborious process which are done only by the professionals. Automating the optic disc detection may help the diagnosis better, faster and reduce the manual process of detection by experts. In automatic optic disc detection, major parts in retina like blood vessels, optic disc and optic cup looks similar which make the optic disc detection more challenging.

Clustering is an unsupervised method which partitions an image into number of clusters. Agglomerative, Divisive, DBSCAN, OPTICS, Fuzzy C Mean and K-means are some of the clustering methods that can be used for segmenting the optic disc. The need for localization of optic disc from fundus image is mandatory as they are mostly used by ophthalmologists for human eye diagnosis [2]. Many machine learning techniques were applied in segmentation and localization of optic disc [3]. Also many researchers have experimented the optic disc detection using Evolutionary approaches [4] and deep learning techniques [5, 6]. Partition-based clustering, Density based clustering and Hierarchical based clustering are few unsupervised learning methods, which were used in optic disc localization of fundus image [7, 8].

The objective of this chapter is to determine and localize the optic disc from the fundus image using appropriate clustering methods like DB Scan clustering, Fuzzy C Mean clustering, Hierarchical clustering and K-Means clustering. These clustering methods were applied on the following dataset namely DRIONS DB dataset, DIARETDB1 dataset and High-Resolution Fundus (HRF) dataset. The suitable dataset for the above clustering methods was determined and the performance was analyzed using various performance measures.

The remainder chapter are organized as, the related work is discussed in Sect. 2, followed by discussion about the clustering methods in Sect. 3, Methodology deployed in detecting optic disc in Sect. 4 and Sect. 5 discuss about the results obtained from the clustering algorithms and Sect. 6 about the conclusion.

2 Related Work

Many researches are been carried out in the field of optic disc segmenting in the fundus image. Table 1 summarizes the various approaches in optic disc segmenting of fundus eye image. Lupascu et al. [9] applied a technique for optic disc detection by identifying the finest corresponding circles. In this technique hundreds of circles were generated by providing the initial points and point of interest. The best fitting circles in the optic disc regions are determined using the regression methods and texture descriptors. Based on optimum properties hundreds of circles were reduced to less than twenty circles and the best optic disc circle is determined by selecting the circle in which the correlation coefficient is maximum. This algorithm was tested on the DRIVE dataset containing 40 images and obtained 95% success rate on optic disc localization and 70% for identifying optic disc. Since the images with low quality were not properly detected this technique could be improved.

Youssif et al. [10] applied a matched filter based algorithm for optic disc segmentation called vessels direction matched filter which uses adaptive histogram and illumination equalization for segmenting the retinal blood vessels. In this algorithm, optic disc vessels direction is roughly matched by using the matched filter and the retinal vessels are segmented using 2-D Gaussian matched filter. The optic disc center candidates are then determined by applying low intensity filters to the segmented retinal vessels. The evaluate of proposed algorithm is done by considering the STARE dataset having 81 images and DRIVE dataset containing 40 images and it is determined that the proposed algorithm has correctly identified optic disc center

Table 1 Methods for optic disc segmentation

Author	Optic disc detection technique	Year	Dataset	No of images
Lupascu et al.	Determine best fit circle through noncollinear points	2008	DRIVE	40
Youssif et al.	Vessel direction matched filter	2008	STARE DRIVE	81 40
Zhu and Rangayyan	Hough Transform	2008	DRIVE STARE	40 81
Lu	Circular transformation	2011	STARE ARIA MESSIDOR	81 120 1200
Welfer et al.	Optic disc location vessels arcade information and optic disc boundary using watershed transform	2010	DRIVE DIARETDB1	40 89
Yin et al.	Hough transform and statistical deformable mode	2011	ORIGA	650
Cheng et al.	PPA elimination	2011	ORIGA	650
Tjandrasa et al.	Hough transform to detect circle and active contours to obtain boundary	2012	DRIVE	40
Fraga et al.	Hough transform and Fuzzy convergence	2012	VARIA	120
Zhang et al.	Characteristics of vessel distribution	2012	STARE DRIVE DIARETDB0 DIARETDB1	81 40 130 89

in 80 fundus images (98.77%) in STARE dataset and 40 images (100%) in DRIVE dataset. The future work of the algorithm aims in applying pre-processing techniques in blood vessel segmentation.

Zhu and Rangayyan [11] applied a method using Hough transform for segmenting the optic disc automatically from fundus image, where the detection of edges is done through sobel operators and canny edge detection followed by detection of circles through Hough transform. Color component of the image was normalized as a pre-processing step and morphological erosion to remove the artifacts. The outliers are removed using median filter. Convolving the pre-processed image with specified operators, vertical gradient and horizontal gradient components of the Sobel operator are attained. Canny operators were applied to determine the edges based on optimization procedures, multiscale analysis and multidirectional derivatives. Once the edges were detected the center and radius of the circle was determined by applying Hough transform. DRIVE dataset and STARE dataset are used in testing the algorithm. The algorithm for sobel method achieved a success rate of 40.24% in STARE dataset and 92.5% in DRIVE dataset and a success rate of 80% in DRIVE dataset and

21.95% in STARE dataset for canny method. By applying additional characteristics of optic disc this algorithm can be improved.

Lu [12] proposed a circular transformation technique to determine and segment the optic disc which detects the color variants and circular boundary. In pre-processing the image intensity is determined from the green component and red components which provides more structural information on optic disc. The image was resized to one third and median filter applied to remove noise on blood vessels. Circular transformation is designed by determined the varying distance within the boundary area points and circular area which reaches minimum when the point lies on centroid region. Maximum variation pixels were filtered and by converting the image optic disc map is determined. The maximum values that are present in the optic disc map will represent the optic disc center. MESSIDOR dataset, ARIA dataset, STARE dataset containing 1200, 120, 81 images respectively are considered for evaluating the proposed technique. The optic disc detection accuracy of this technique was found to be 98.77% for MESSIDOR dataset, 99.75 for STARE dataset and 97.5% for ARIA dataset and the accuracy of optic disc segmentation of 93.4% for STARE dataset and 91.7% for ARIA datasets.

Welfer et al. [13] applied an adaptive method to optic disc segmenting the using mathematical morphology, which is robust for the images acquired from varying illumination and acquisition conditions. This methodology has two stages of optic disc detection. Initially the location of optic disc is detected from the vessels arcade information by determining the background and foreground of green channel image using RMIN operator. Next stage the boundary of optic disc is determined using watershed transform. The evaluation was done on DRIVE data set containing 40 images and DIARETDB1 dataset containing 89 images. It achieves a success rate in optic disc localization of DIARETDB1 dataset is 97.75% and DRIVE dataset is 100%. The future works of this work shall consider retina structures like fovea for optic disc detection.

Yin et al. [14] applied a technique which uses circular statistical deformable model, Hough transform and edge detection to determine the optic disc. In pre-processing the effect of blood vessels are reduced by carrying out voting scheme-based heuristics for selection of optimal channel. The optic disc center is determined using Hough transform, while the optic disc boundary is determined using statistical deformable model. Subsequently optic disc circle is approximated using Hough transform and the optic disc center and diameter are determined. Finally, boundaries of image texture are fine-tuned by applying statistical deformable model and optic disc boundary is smoothed by direct least squared ellipse fitting method. The evaluation of the applied technique was determined using ORIGA dataset having 650 images. The result states that this technique have 11.3% average error in the overlapping area and 10.8% average absolute area error.

Cheng et al. [15] applied an elimination method called peripapillary atrophy (PPA) elimination for optic disc segmentation. This proposed method has three stages of implementation include edge filtering, Hough transform and β-PPA detection. In the initial stage the noise is removed by applying low pass filter and the optic disc edges and region of interest are extracted. In edge filtering α and β -PPA was eliminated.

α-PPA includes the epithelial cells which are duskier than the optic disc and β-PPA is similar to the color of optic disc with region of chorioretinal atrophy. α-PPA is determined by comparing the threshold value with the region of interest. Hough transform is applied as the second stage of PPA elimination. In the final stage PPA elimination is to detect β-PPA, due to its similar color as optic disc make the detection difficult when compared to α-PPA. The ring area of optic disc boundary was determined and was segmented into quarters. The feature point is detected by considering the local maxima and minima. The feature points are compared with the threshold, If the feature points are higher than the threshold means than β-PPA is present. The edge points are removed which present along the disc boundary that is detected. The new optic disc boundary is determined by reapplying the constrained elliptical Hough transform. The proposed method is evaluating by considering ORIGA dataset consisting of fundus images. The results infer that the method has 10% average overlapping error, 7.4% average absolute area error and 4.9% average error of vertical disc diameter.

Tjandrasa et al. [16] applied a technique that is initiated by converting the image into grayscale followed by image enhancement using homomorphic filtering to remove the uneven illumination, once the image is enhanced the blood vessels are removed to simplify segmentation task. The low pixel values are determined and the blood vessels are blurred using median filter. The next stage in optic disc segmentation is to perform Hough transform to determine the matched optic disc location. Subsequent optic disc boundaries are determined by performing active contour model. The DRIVE dataset is used for evaluation the experimental results and it found that the optic nerve is segmented with an accuracy of 75.56%.

Fraga et al. [17] applied a model with faster localization and segmentation. This technique uses multiscale retinx algorithm for pre-processing. Fuzzy Hough transform and fuzzy convergence of the blood vessels are used to identify the optic disc location. Optic disc segmentation is then carried out by applying canny filters. Finally, the accuracy of segmentation was determined using histogram information. VARIA dataset containing 120 images is considered for evaluation. The result found that the success rate of proposed algorithm is 100% for optic disc localization and 93.36% for optic disc segmentation.

Zhang et al. [18] used a technique in which the localization is done through 1D projection to determine the optic disc with its appearance characteristic of optic disc in vertical location and vessel distribution of optic disc in horizontal location. The horizontal location of optic disc is determined using vascular degree and with the edge gradients around the optic disc, the vertical location was obtained. In preprocessing the region of interest was identified using mask operation. Once the preprocessing is complete optic disc horizontal location is computed based on vascular degree, then the vascular scatter degree is determined by defining vertical window and sliding over the vessel map. The optic disc vertical location is computed using 1D vertical projection signal which is obtained by defining the rectangular window centered at optic disc horizontal location and sliding over Gabor filter map. Evaluation of proposed algorithm was done by considering the DRIVE dataset containing 40 images, STARE dataset containing 81 images, DIARETDB0 dataset containing 130 images and DIARETDB1 dataset containing 89 images. The outcomes infer that the success rate of DRIVE dataset is 100%, STARE dataset is 91.4%, DIARETDB0 dataset is 95.5% and DIARETDB1 dataset is 92.1%.

3 Clustering Methods

Clustering is broadly used in the area of image segmentation. This clustering methods forms a variety of separate clusters by partitioning the images [19]. The clustering is generally classified as Density based clustering method, Hierarchical based clustering method and Partition based clustering method. Figure 2 represent the complete classification of clustering methods.

Partition method:

It divides the data into number of partitions based on their characteristics and similarity. Once the user specifies the number of partitions to be made the algorithm divides the data into as many partitions, where every partition represents a cluster. Partition clustering methods may include K-Means, K-medoids, CLARA, FCM, etc.

Hierarchical method:

It involves formation of clusters in a planned order, which will be grouping into tree of clusters. Initially it will identify and merge the two comparable clusters which are close to each other and will be continued until all the comparable clusters are combined each other. The major aim of hierarchical clustering method is to generate a hierarchical series of nested cluster. The basic methods of hierarchical clustering include Agglomerative and Divisive clustering methods. Agglomerative clustering

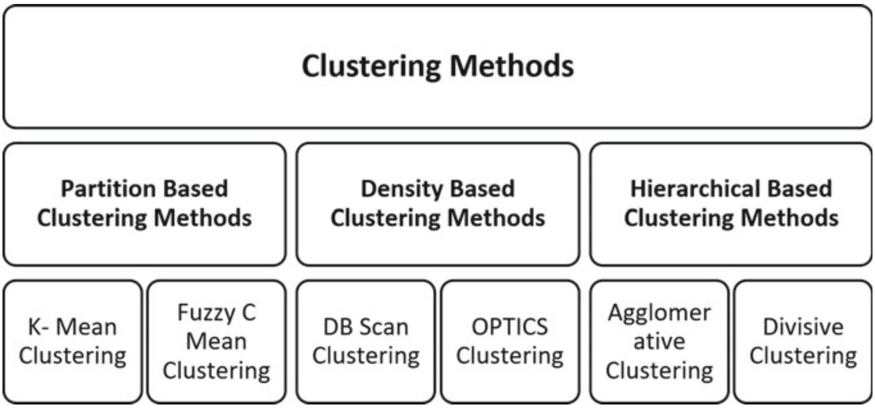


Fig. 2 Classification of clustering methods

method uses an approach which is bottom up while the Divisive which uses the top down approach.

Density based method:

It is a clustering method where the clusters are determined by forming a distinctive cluster one for high density data and another for low density data. Typically, the low-density data are considered as outlier or noise. The different types of density-based methods may include DBScan, Optics, etc.

3.1 Agglomerative Clustering

Agglomerative clustering is a bottom up method of hierarchical clustering in which individual clusters are considered form initial data point. The comparable clusters are joined to other clusters in each iteration until the formation of a cluster.

Agglomerative clustering Algorithm
Determine Matrix
Let the unique Cluster be a Data Point
Repeat until formation of N clusters
  Merge the Two Neighboring Clusters.
  Update Matrix.

3.2 K-Means Clustering

Most extensively used iterative techniques are K Means clustering that partition the data into discrete non-overlapping clusters with individual data fit into the group. It ensures individual cluster are different by making similar inter-cluster data points. Sum of squared distance between data points are assigned as a cluster to data points [20].

K-Means clustering Algorithm
Input No of cluster (N)
Initialize Centroids by shuffling the dataset
Repeat until centroid does not have any change
  Compute data points sum of squared distance
  Assign data point to each of the nearby cluster
  Determine mean and compute cluster centroids.

3.3 Fuzzy C Mean Clustering

It is an iterative clustering technique where as in fuzzy logic, each point has a grade off fitting to a cluster, rather entirely fitting to a single cluster. Thus, the degree of points on the centre of the cluster will be superior the points on the edge of a cluster. Fuzzy c-means tries partition of a partial group of elements into group of fuzzy clusters with specified condition. It is built on minimization of succeeding objective function [21].

$$\mathrm{J(X, Y)} = \sum_{i=1}^{n}\sum_{q=1}^{a}\left(X_{p,q}\right)^{r}\left|\left|w_{p} - y_{q}\right|\right|^{2}$$

Here,

- r—Real nos > 1.
- a—No of clusters.
- n—No of data.
- $\| w_p - y_q \|$—distance between data point p and present cluster centre q.
- X_{pq}—Membership value.

Fuzzy C Mean Clustering Algorithm
Each cluster are assigned with a random centroid
Repeat until centroid does not have any change.
Determine centroid based on the distance between the cluster center and datapoint.
Centroid is **updated** based on new member function
If new and original centroid distance is lower than threshold
then stop because it is an Epsilon.
else
Continue until the condition becomes true.

3.4 DBSCAN Clustering

Most common density-based clustering is a DBSCAN method. It determines samples of high density and low density, then develops the cluster with samples of high density that are more suitable for data with similar density [22].

DBSCAN Clustering Algorithm
Select a point randomly (P)

Identify all the point (P) reachable Point Density
Continue till the points are processed
If Central Point are P
Cluster are formed.
else boundary point is P
density reachable points are Zero.

4 Methodology

The methodology deployed to detect the optic disc is represented in the block diagram below Fig. 3. The fundus eye image is provided as an input which is resized to reduce computational power and preprocessed by selecting the green color band and smoothing. Once the preprocessing is complete clustering technique is applied to it and optic disc region are masked and the portion of optic disc alone is segmented.

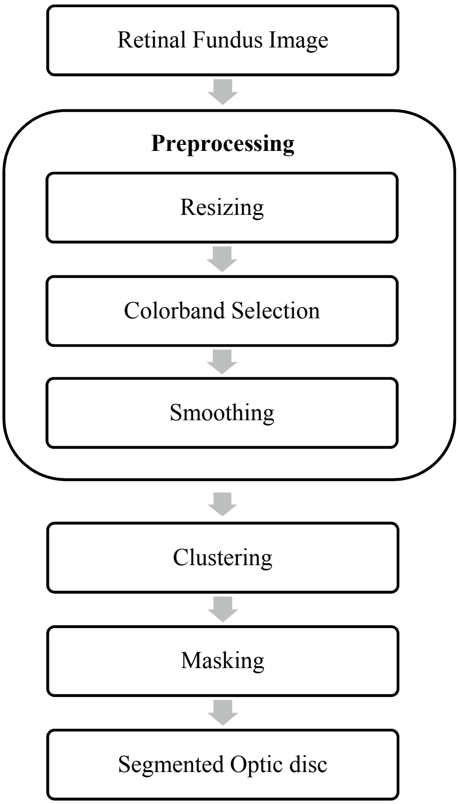


Fig. 3 Process to detecting optic disc

4.1 Retinal Image Pre-processing

The fundus images are given as an input for optic disc detection. The fundus images shall contain different type of noise and may consume more computational capability. To effectively handle those noises and process it effectively the images need to be enhanced by the pre-processing techniques like Resizing, Color band selection and Smoothing [23].

4.1.1 Resizing

The fundus images may be of varying sizes and shall require enormous computational facility and sometimes the images may be too large to fit onto the screen. Thus, the original image is be resized as per the requirement without removing or altering any of its properties. The fundus images in all the dataset are resized to width of 350 and height of 233 (i.e., 350×233).

4.1.2 Color Band Selection

Every pixel is represented either using Red, Green or Blue color with their numeric values ranging from 0 to 255. The colored images may require high computational capability and sometimes does not have pathological structure facts. So, the processing of color image is not a good idea. To overcome this color bands can be separated from the fundus image. The color band which have good contrast is selected for processing the optic disc. In this work it's found that green color band is good and more suitable than blue and red color band.

4.1.3 Smoothing

The green color band fundus image obtained shall comprises of noises and blood vessels. These noises are removed by applying a morphological operator called dilation, which will make improve the images visibility and even fills the small holes that are present in the image then, the high frequency components are removed and the image is made ready for processing by applying gaussian blur operations.

4.1.4 Clustering Methods to Segment Optic Disc

Once the pre-processing of the fundus image is complete, the clustering techniques are applied to determine the optic disc in the fundus image. Clustering method forms a group of clusters by dividing the image into pixels based on their similarities. The region of optic disc will be having the similar property, so they lie on the single cluster. The optic disc is then segmented from the formed cluster.

4.1.5 Masking

The brightest pixels in the clustered image contains the optic disc. Masking the brightest pixels is needed to determine the optic disc region while the other regions are masked out. A circle of fixed radius is drawn by determining the centroid. The pictorial representation of optic disc segmentation is shown in the Fig. 4.

5 Results and Discussion

5.1 Dataset Description

The algorithm is realized using Open Source Computer Vision Library and is implemented on Intel (R) core i3 processor with a RAM of 4 Giga Bytes. The algorithm was verified with the following publicly available databases like DIARETDB1 dataset,

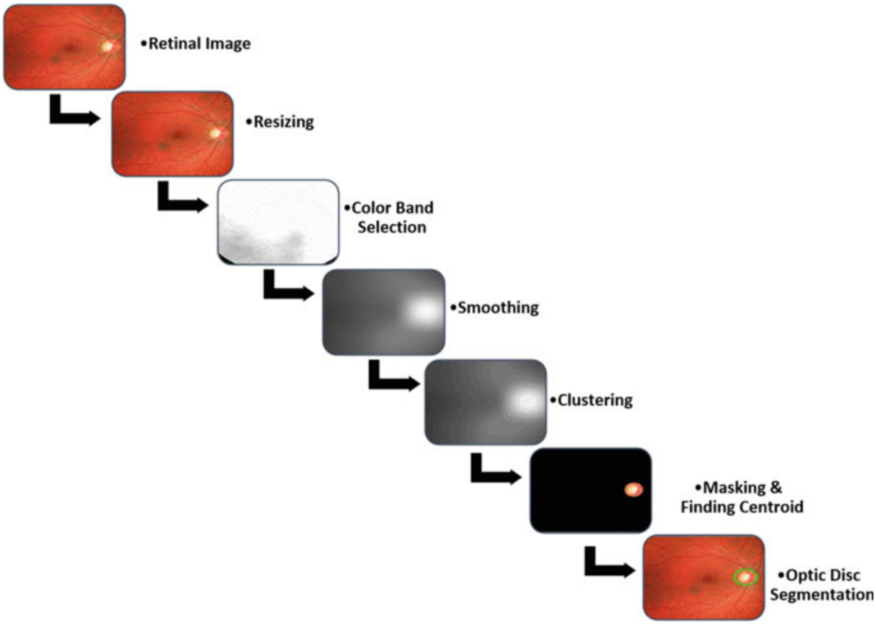


Fig. 4 Pictorial representation of optic disc segmentation in fundus eye image using clustering

Table 2 Dataset description

Dataset	Total images	Healthy images	Unhealthy images	Camera and view of fundus image
HRF	45	15	30	CR-1 fundus camera by canon with 45° view
DIARETDB1	89	05	84	Digital fundus camera with 50° view
DRIONS	110	06	104	Colour analogical fundus camera

DRIONS dataset and High-Resolution Fundus dataset. The detailed description of the three datasets are show in the below Table 2.

5.1.1 High Resolution Fundus Dataset

High Resolution Fundus dataset are available to public usage, which consist of 45 images with an equal spread of 15 fundus images of patients who are healthy, 15 fundus images of patients having diabetic retinopathy and 15 fundus images of patients having glaucomatous. These images are shot on Canon CR-1 fundus camera with 45° view with a resolution of 3304 × 2336 [24].

5.1.2 DIARETDB1 Dataset

DIARETDB1 dataset consist of 89 images with 5 images are considered healthy images and 84 images encompasses mild non-proliferative symptoms considered as unhealthy images. The digital fundus camera was used to shot images with 50° view with a resolution of 1500 × 1152 [25].

5.1.3 DRIONS-DB Dataset

DRIONS-DB is a publicly available dataset consist of 110 images with 6 Healthy images and 104 unhealthy images. These images in this dataset are captured on color analogical fundus camera with a resolution of 600 × 400. [26].

5.2 Result Analysis

Clustering methods K-means clustering, Agglomerative clustering, DBSCAN clustering and Fuzzy C-means clustering techniques performance and accuracy determined and compared on the datasets like HRF, DIARETDB1 and DRIONS datasets.

The optic disc localization by K Means clustering in Figs. 5 and 6, Fuzzy C Means clustering in Figs. 7 and 8, Agglomerative clustering in Figs. 9 and 10 and DBSCAN clustering in Figs. 11 and 12 are shown below. The detected optic disc portion are represented by a green circle.

Figure 13 shows the various stages of optic disc segmenting in fundus image on HRF dataset, DIARETDB1 dataset and DRIONSDB dataset, where the fundus images are given as an input which is been pre-processed by resizing, green channel selection and smoothening. Then the clustering techniques is applied to the processed image by determining and segmenting the optic disc region with a circle.

Figure 14 shows the optic disc segmentation by DBSCAN clustering, Hierarchical clustering, FCM clustering and K Means clustering methods on publicly available databases HRF dataset. The circle in the image represent the optic disc region that is been fittingly segmented by the clustering techniques.

The optic disc segmentation is done by making a circle on the optic disc region on DIARETDB1 dataset by DBSCAN clustering, Hierarchical clustering, FCM clustering and K Means clustering methods is shown in the below Fig. 15.

Fig. 5 Input fundus image for DBSCAN clustering

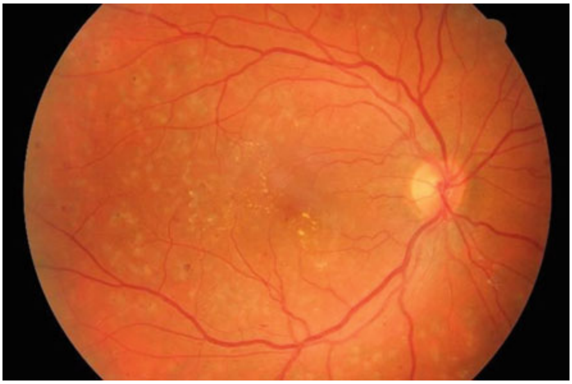
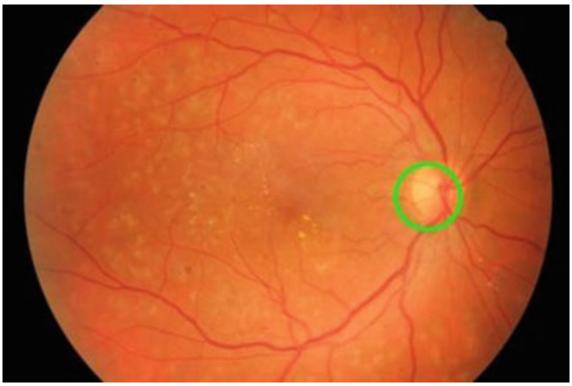


Fig. 6 Localized optic disc by DBSCAN clustering



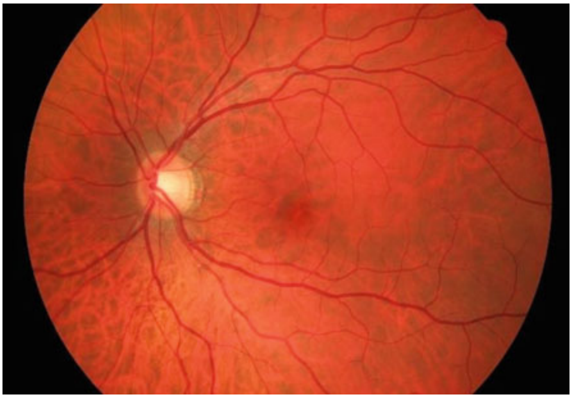


Fig. 7 Input fundus image for Hierarchical clustering

Fig. 8 Localized optic disc by Hierarchical clustering

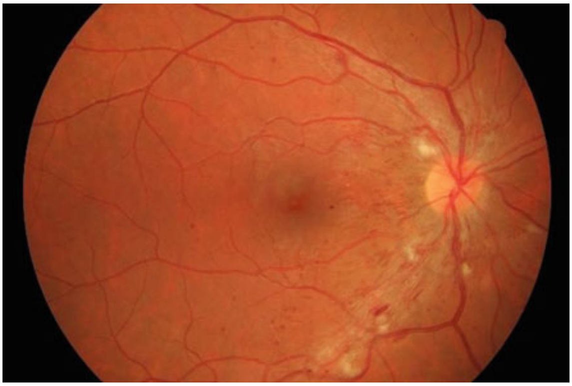


Fig. 9 Input fundus image for FCM clustering

Fig. 10 Localized optic disc by FCM clustering

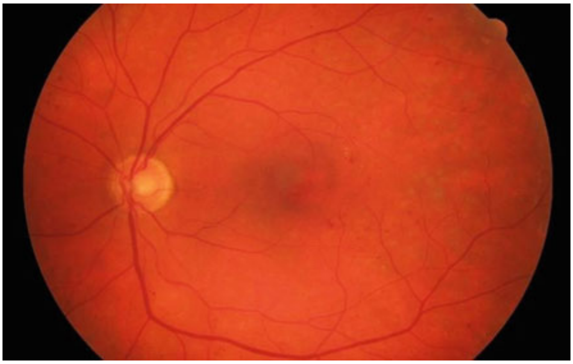


Fig. 11 Input fundus image for K-Means clustering

Fig. 12 Localized optic disc by K-Means clustering

Figure 16 shows the sample images of DRIONS dataset where the optic disc is detection and segmentation using DBSCAN, Hierarchical, FCM and K Means clustering techniques. The sample images are taken randomly and be evaluated and displayed in the below figure.

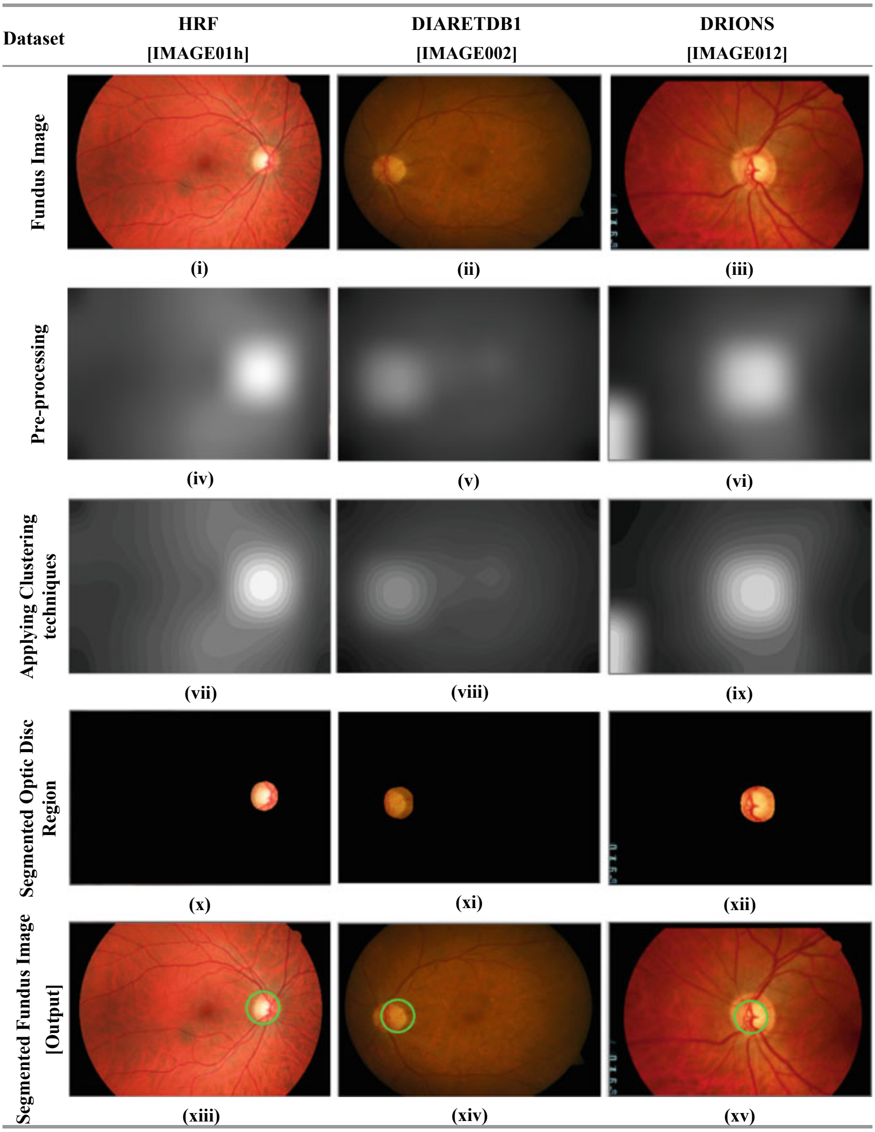


Fig. 13 Stages in segmenting optic disc from fundus image. (i) Input fundus image of HRF dataset, (ii) input fundus image of DIARETDB1 dataset, (iii) input fundus image of DRIONS dataset, (iv) pre-processing of HRF dataset, (v) pre-processing of DIARETDB1 dataset, (vi) pre-processing of DRIONS dataset, (vii) applying clustering technique for HRF dataset, (viii) applying clustering technique for DIARETDB1 dataset, (ix) applying clustering technique for DRIONS dataset, (x) segmented optic disc region of HRF dataset, (xii) segmented optic disc region of DIARETDB1 dataset, (xii) segmented optic disc region of DRIONS dataset, (xiii) output of fundus image in HRF dataset, (xiv) output of fundus image in DIARETDB1 dataset, (xv) output of fundus image in DRIONS dataset

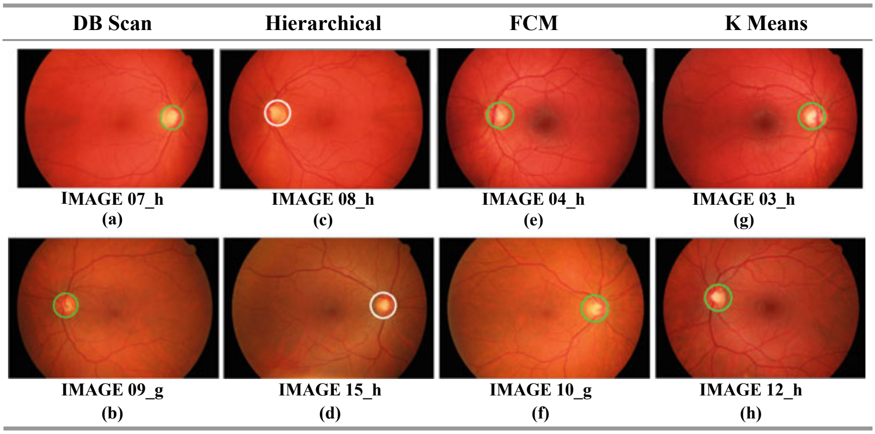


Fig. 14 Examples of correct detection of optic disc on HRF dataset circle represent the segmented region of optic disc. **a** and **b** DBSCAN clustering on HRF dataset, **c** and **d** hierarchical clustering on HRF dataset, **e** and **f** FCM clustering on HRF dataset, **g** and **h** K Means clustering on HRF dataset

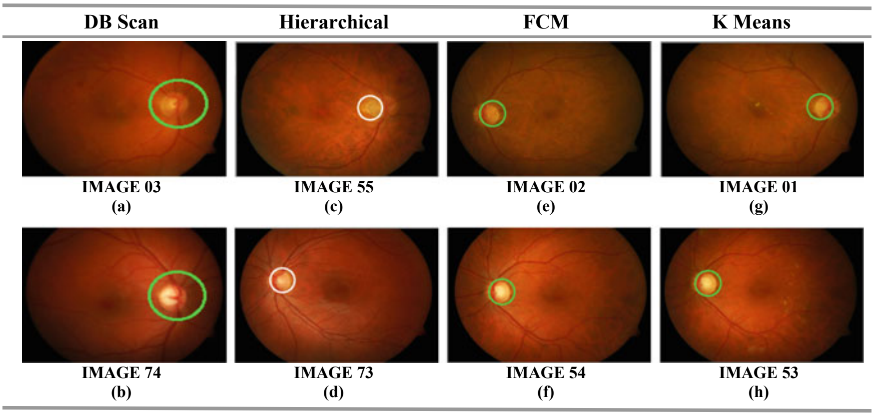


Fig. 15 Examples of correct detection of optic disc on DIARET DB1 dataset circle represent the segmented region of optic disc. **a** and **b** DBSCAN clustering on DIARETDB1 dataset, **c** and **d** hierarchical clustering on DIARETDB1dataset, **e** and **f** FCM clustering on DIARETDB1dataset, **g** and **h** K Means clustering on DIARETDB1dataset

Optic disc segmentation in HRF, DIARETDB1 and DRIONS dataset were done by K Means, Fuzzy C Mean, Hierarchical and DBSCAN clustering techniques. Yet segmentation was non proper in few images because of the image quality. Figure 17 shows some difficult cases in each dataset and in each clustering techniques where the segmentation of optic disc failed.

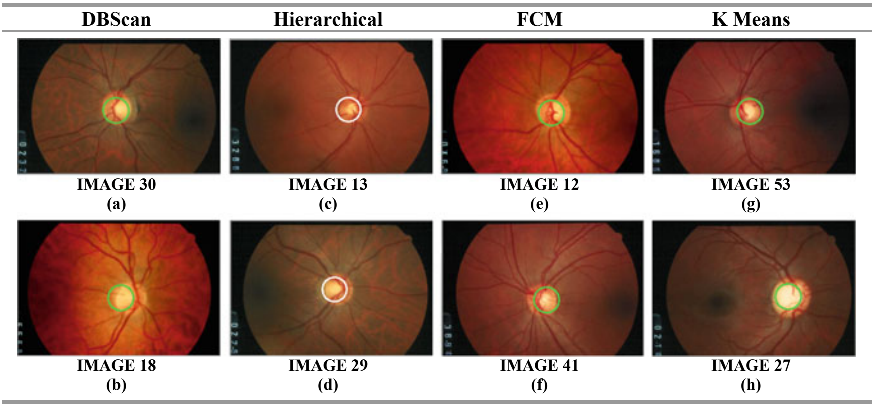


Fig. 16 Examples of correct detection of optic disc on DRIONS dataset circle represent the segmented region of optic disc. **a** and **b** DBSCAN clustering on DRIONS dataset, **c** and **d** hierarchical clustering on DRIONS dataset, **e** and **f** FCM clustering on DRIONS dataset, **g** and **h** K Means clustering on DRIONS dataset

5.3 *Optic Disc Detection*

The region of optic disc in the fundus image are determined by comparing ground truth centre of the optic disc and manually determined optic disc centre. The detection of optic disc is said to be done (i.e., Hit) if the below equation [19] is fulfilled [4].

$$\sqrt{(\boldsymbol{A_m} - \boldsymbol{A}_d)^2 + (\boldsymbol{B_m} - \boldsymbol{B_d})^2} < \boldsymbol{X_{mean}}$$

Here, A_m and B_m is manually labelled, A_d and B_d are detected coordinates, X mean denotes the datasets standard average radii. The process is said to be Hit if the equation is satisfied and vice versa if it is a Miss. Table 3 represent the Hit rate determined by applying the clustering algorithms. Figure 18a, b shows the Graphical representation of the clustering techniques on datasets for 10 and 20 iterations.

5.4 *Performance Evaluation*

The performance measure of detecting optic disc in fundus eye image are determined using various parameters like Positive Predicted Value, False Discovery Rate, Accuracy, F1 Score, Error Rate, Precision and Recall [27], the description of the measure is shown in Table 4. The outcome values of segmented fundus images are determined by evaluating True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) as shown in Fig. 19.

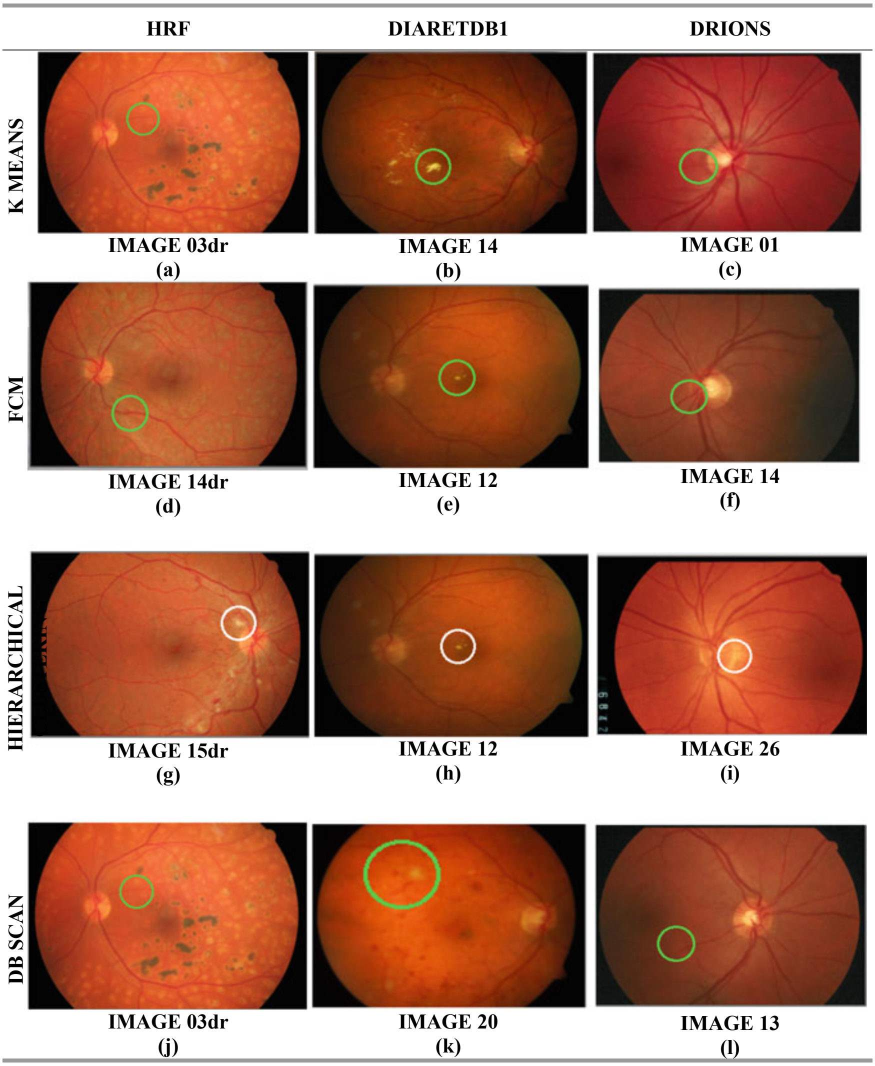


Fig. 17 Incorrect optic disc segmentation in difficult cases. **a** K Means clustering on HRF dataset, **b** K Means clustering on DIARETDB1 dataset, **c** K Means clustering on DRIONS dataset, **d** FCM clustering on HRF dataset, **e** FCM clustering on DIARETDB1 dataset, **f** FCM clustering on DRIONS dataset, **g** hierarchical clustering on HRF dataset, **h** hierarchical clustering on DIARETDB1 dataset, **i** hierarchical clustering on DRIONS dataset, **j** DBSCAN clustering on HRF dataset, **k** DBSCAN clustering on DIARETDB1 dataset, **l** DBSCAN clustering on DRIONS dataset

Table 3 Clustering techniques accuracy

			FCM	DIARETDB1	DRIONS
	No of iteration	**Total images**	45	89	110
K Means	10	Optic disc detection	45	68	82
		Optic disc missed	0	21	28
		Hit rate	**100%**	**76%**	**75%**
	20	Optic disc detection	45	68	85
		Optic disc missed	0	21	25
		Hit rate	**100%**	**76%**	**77%**
FCM	10	Optic disc detection	45	68	78
		Optic disc missed	0	21	32
		Hit rate	**100%**	**76%**	**71%**
	20	Optic disc detection	45	68	74
		Optic disc missed	0	21	36
		Hit rate	**100%**	**76%**	**67%**
Hierarchical	10	Optic disc detection	45	67	110
		Optic disc missed	0	22	0
		Hit rate	**100%**	**75%**	**100%**
	20	Optic disc detection	45	68	110
		Optic disc missed	0	21	0
		Hit rate	**100%**	**76%**	**100%**
DB scan	10	Optic disc detection	45	78	75
		Optic disc missed	0	32	35
		Hit rate	**100%**	**71%**	**68%**
	20	Optic disc detection	45	74	76
		Optic disc missed	0	36	34
		Hit rate	**100%**	**67%**	**69%**

5.5 Discussion

The performance parameters like Positive Predicted Value, False Discovery Rate, Accuracy, Error Rate, Precision and Recall are determined by applying clustering algorithms and identifying True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (FN). The observed result for 10 and 20 iterations are tabulated in the below Table 5 and 6 respectively.

The performance parameters like Positive Predicted Value, False Discovery Rate, Accuracy, Error Rate, Precision and Recall are determined by analyzing clustering techniques for 10 and 20 iterations on HRF dataset and the average values obtained by the techniques on the various performance measures is shown in the below Table 7.

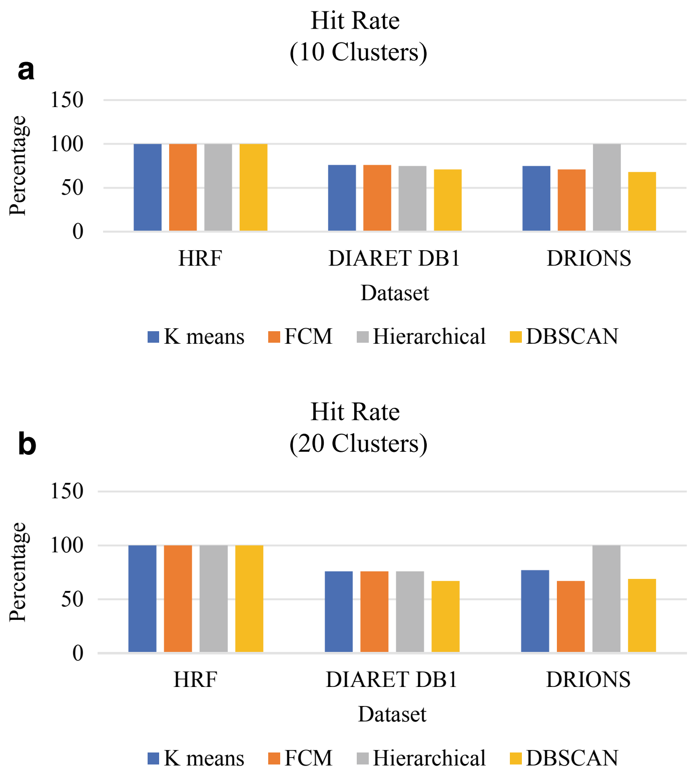


Fig. 18 **a** Hit rate for 10 clusters. **b** Hit rate for 20 clusters

Table 4 Description of performance measure

Measure	Description
Positive predicted value (PPV)	(True Positive)/(False Positive + True Positive)
False discovery rate (FDR)	(False Positive)/(True Positive + False Positive)
Accuracy	(True Positive + True Negative)/Total
Error rate	(False Negative + False Positive)/(False Negative + False Positive + True Negative + True Positive)
F1 score	2 * (Precision * Recall/(Precision + Recall))
Precision	(True Positive)/(True Positive + False Positive)
Recall	(True Positive)/(True Positive + False Negative)

Where,
True Positive (TP) = Correctly detected Optic disc
False Negative (FN) = Unable to detect Optic disc
False Positive (FP) = Optic disc detect but it's incorrectly detected
True Negative (TN) = No Optic Disc No Detection

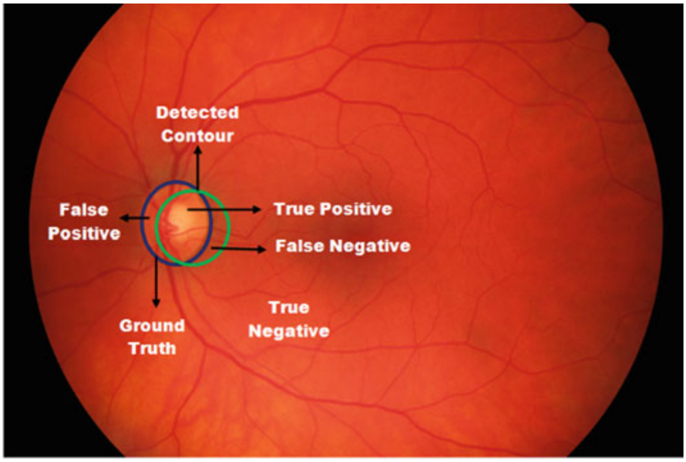


Fig. 19 Determining TP, TN, FP, FN in fundus image

Table 5 Performance measure analysis for 10 iterations

	Dataset	Total images	FP	TP	TN	FN
K means	HRF	45	4	41	0	0
	DIARET DB 1	89	9	59	0	21
	DRIONS	110	8	74	0	28
FCM	HRF	45	1	44	0	0
	DIARET DB 1	89	9	59	0	21
	DRIONS	110	8	70	0	32
Hierarchical	HRF	45	1	44	0	0
	DIARET DB 1	89	5	62	0	22
	DRIONS	110	5	105	0	0
DBSCAN	HRF	45	3	42	0	0
	DIARET DB 1	89	2	42	0	45
	DRIONS	110	1	74	0	35

The performance of the clustering techniques is determined by analysing clustering technique for 10 and 20 iterations on DIARETDB1 dataset and the average values obtained by the techniques on the various performance measures is shown in the below Table 8.

Table 9 represents the result obtained on the DRIONS datasets for 10 iterations and 20 iterations along with the average iteration values by the clustering techniques. Similar to the above results hierarchical clustering technique outperforms others in most of the cases.

Table 6 Performance measure analysis for 20 iterations

	Dataset	Total Images	FP	TP	TN	FN
K Means	HRF	45	2	43	0	0
	DIARET DB 1	89	8	60	0	21
	DRIONS	110	9	76	0	25
FCM	HRF	45	3	42	0	0
	DIARET DB 1	89	8	60	0	21
	DRIONS	110	8	66	0	36
Hierarchical	HRF	45	1	44	0	0
	DIARET DB 1	89	5	63	0	21
	DRIONS	110	4	106	0	0
DBSCAN	HRF	45	4	41	0	0
	DIARET DB 1	89	2	42	0	45
	DRIONS	110	1	75	0	34

Table 7 HRF dataset on clustering techniques

		DB scan clustering	Hierarchical clustering	FCM clustering	K-mean clustering
Precision	Iteration 10	93.33	97.78	97.78	91.11
	Iteration 20	91.11	97.78	93.33	95.56
	Average	92.22	97.78	95.56	93.33
Recall	Iteration 10	100.00	100.00	100.00	100.00
	Iteration 20	100.00	100.00	100.00	100.00
	Average	100.00	100.00	100.00	100.00
Accuracy	Iteration 10	93.33	97.78	97.78	91.11
	Iteration 20	91.11	97.78	93.33	95.56
	Average	92.22	97.78	95.56	93.33
F1 score	Iteration 10	96.55	98.88	98.88	95.35
	Iteration 20	95.35	98.88	96.55	97.73
	Average	95.95	98.88	97.71	96.54
Error rate	Iteration 10	6.67	2.22	2.22	8.89
	Iteration 20	8.89	2.22	6.67	4.44
	Average	7.78	2.22	4.44	6.67
Positive predicted value (PPV)	Iteration 10	0.93	0.98	0.98	0.91
	Iteration 20	0.91	0.98	0.93	0.96
	Average	0.92	0.98	0.96	0.93
False discovery rate (FDR)	Iteration 10	0.07	0.02	0.02	0.09
	Iteration 20	0.09	0.02	0.07	0.04
	Average	0.08	0.02	0.04	0.07

Table 8 DIARETDB1 dataset on clustering techniques

		DB scan clustering	Hierarchical clustering	FCM clustering	K-mean clustering
Precision	Iteration 10	95.45	92.54	86.76	86.76
	Iteration 20	95.45	92.65	88.24	88.24
	Average	95.45	92.59	87.50	87.50
Recall	Iteration 10	48.28	73.81	73.75	73.75
	Iteration 20	48.28	75.00	74.07	74.07
	Average	48.28	74.40	73.91	73.91
Accuracy	Iteration 10	47.19	69.66	66.29	66.29
	Iteration 20	47.19	70.79	67.42	67.42
	Average	47.19	70.22	66.85	66.85
F1 score	Iteration 10	64.12	82.12	79.73	79.73
	Iteration 20	64.12	82.89	80.54	80.54
	Average	64.12	82.51	80.13	80.13
Error rate	Iteration 10	52.81	30.34	33.71	33.71
	Iteration 20	52.81	29.21	32.58	32.58
	Average	52.81	29.78	33.15	33.15
Positive predicted value (PPV)	Iteration 10	0.95	0.93	0.87	0.87
	Iteration 20	0.95	0.93	0.88	0.88
	Average	0.95	0.93	0.88	0.88
False discovery rate (FDR)	Iteration 10	0.05	0.07	0.13	0.13
	Iteration 20	0.05	0.07	0.12	0.12
	Average	0.05	0.07	0.13	0.13

The K-Means clustering method was applied on the DRIONS, DIARETDB1 and HRF dataset for 10 iterations and the results are tabulated in Fig. 20. It is observed from the results that the accuracy of K-Means clustering method on HRF dataset is 91%, which is greater than the other two DIARETDB1 and DRIONS dataset which are have an accuracy of 66% and 67% respectively. Similarly, the K-Means clustering was applied on the DRIONS, DIARETDB1 and HRF dataset for 20 iterations and the obtained results are tabulated in Fig. 21. The result infers that the accuracy of HRF dataset is much better than the other dataset set with HRF dataset having an accuracy of 96% compared to 67% and 69% of DIARETDB1 and DRIONS dataset respectively.

The Performance of hierarchical clustering was determined by applying it on DRIONS, DIARETDB1 and HRF dataset for 10 iteration and the results are tabulated in Fig. 22. With the observation from the results it is inferred that accuracy of hierarchical clustering on HRF dataset is 98% which is greater when compared to the other dataset DIARETDB1 and DRIONS having an accuracy of 70% and 95% respectively. Similarly, Fig. 23 represent the results obtained by performing hierar-

Table 9 DRIONS dataset on clustering techniques

		DB scan clustering	Hierarchical clustering	FCM clustering	K-mean clustering
Precision	Iteration 10	98.67	95.45	89.74	90.24
	Iteration 20	98.68	96.36	89.19	89.41
	Average	98.68	95.91	89.47	89.83
Recall	Iteration 10	67.89	100.00	68.63	72.55
	Iteration 20	68.81	100.00	64.71	75.25
	Average	68.35	100.00	66.67	73.90
Accuracy	Iteration 10	67.27	95.45	63.64	67.27
	Iteration 20	68.18	96.36	60.00	69.09
	Average	67.73	95.91	61.82	68.18
F1 score	Iteration 10	80.43	97.67	77.78	80.43
	Iteration 20	81.08	98.15	75.00	81.72
	Average	80.76	97.91	76.39	81.08
Error rate	Iteration 10	32.73	4.55	36.36	32.73
	Iteration 20	31.82	3.64	40.00	30.91
	Average	32.27	4.09	38.18	31.82
Positive predicted value (PPV)	Iteration 10	0.99	0.95	0.90	0.90
	Iteration 20	0.99	0.96	0.89	0.89
	Average	0.99	0.96	0.89	0.90
False discovery rate (FDR)	Iteration 10	0.01	0.05	0.10	0.10
	Iteration 20	0.01	0.04	0.11	0.11
	Average	0.01	0.04	0.11	0.10

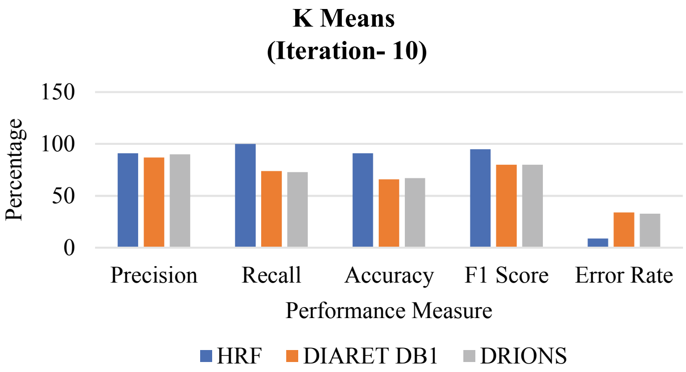


Fig. 20 Performance of K-Means clustering techniques with 10 iterations

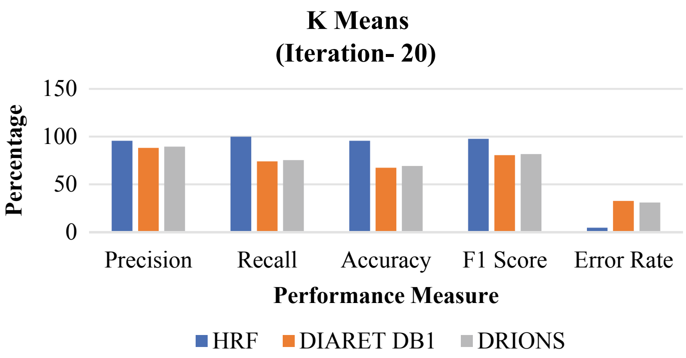


Fig. 21 Performance of K-Means clustering techniques with 20 iterations

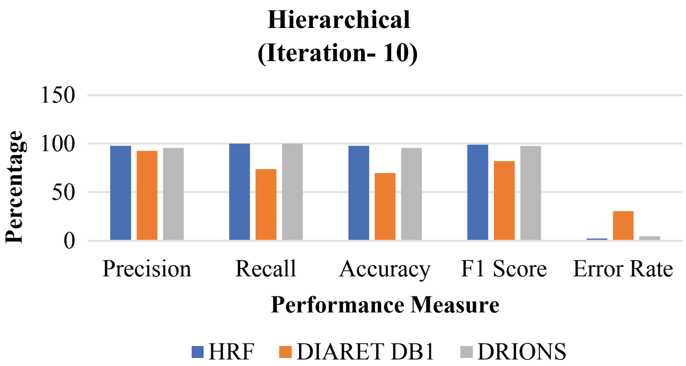


Fig. 22 Performance of hierarchical clustering techniques with 10 iterations

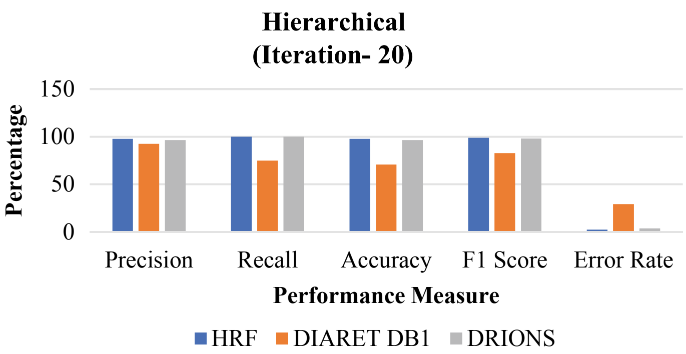


Fig. 23 Performance of hierarchical clustering techniques with 20 iterations

chical clustering on dataset for 20 iterations and the result infer that 98% accuracy on HRF dataset is better than 95% accuracy on DRIONS dataset while the DIARET DB1 dataset have an accuracy of 70% which is less compared to the others datasets.

Fuzzy C Mean clustering was applied on DRIONS, DIARETDB1 and HRF dataset for 10 iteration and 20 iterations respectively. The results obtained from FCM clustering method with 10 iterations is represented in Fig. 24 and for 20 iterations in Fig. 25. The results infer that the performance of HRF data set in both 10 and 20 iterations are better than the other two datasets. The accuracy of HRF dataset for 10 iteration is 98% with comparison to DIARETDB1 and DRIONS with 66% and 64% respectively. While for 20 iterations the 93% accuracy of HRF dataset is better when compared to 67% for DIARETDB1 and 60% for DRIONS dataset.

The Performance of DBSCAN clustering method was determined by applying it on DRIONS, DIARETDB1 and HRF dataset for 10 and 20 iterations. Figure 26 represent the result obtained from DBSCAN clustering method for 10 iterations. It is observed that the accuracy of HRF dataset is 93% which is way better when

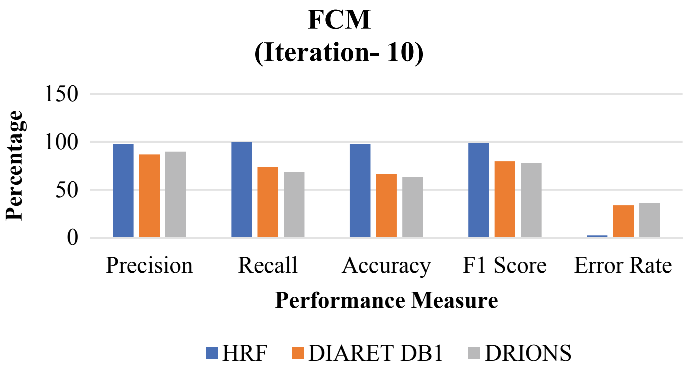


Fig. 24 Performance of FCM clustering techniques with 10 iterations

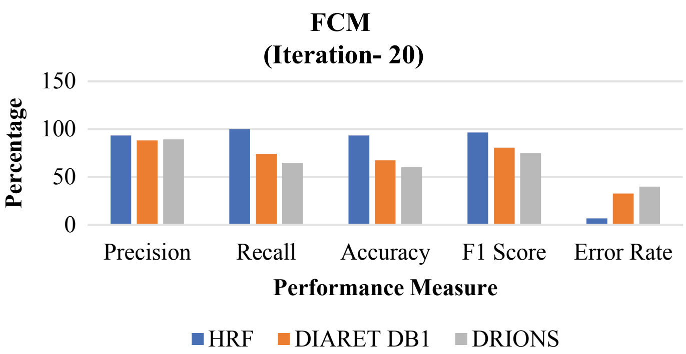


Fig. 25 Performance of FCM clustering techniques with 20 iterations

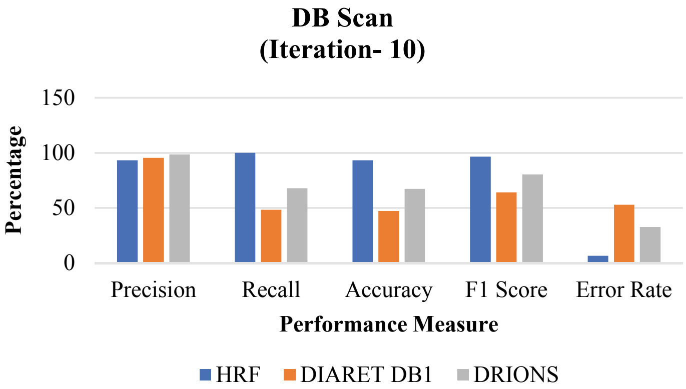


Fig. 26 Performance of DBSCAN clustering techniques with 10 iterations

compared to other DIARETDB1 dataset with 47% and DRIONS dataset with 67%. Similarly, Fig. 27 represent the results obtained by performing DBSCAN clustering on dataset for 20 iterations and the result infer that 91% accuracy on HRF dataset is better than 47% accuracy on DIARETDB1 dataset and 68% accuracy on DRIONS dataset.

The positive predicted values of the clustering algorithms on all the three datasets are determined and is tabulated below. Figures 28 and 29 illustrate that the probability of positively predicted values of hierarchical clustering is better HRF and DIARETDB1 database, while DBSCAN clustering is better in DRIONS dataset for 10 iterations. For 20 iterations the positively predicted values of hierarchical clustering are greater for HRF dataset while DBSCAN clustering is better for DIARETDB1 and DRIONS datasets.

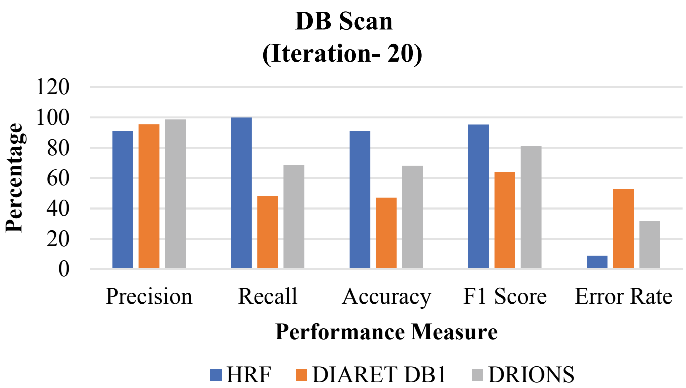


Fig. 27 Performance of DBSCAN clustering techniques with 20 iterations

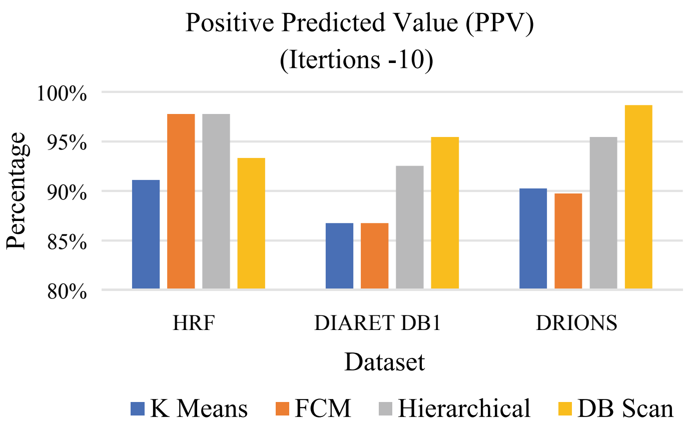


Fig. 28 Positive predicted values of clustering techniques on DRIONS, DIARETDB1 and HRF dataset for 10 iterations

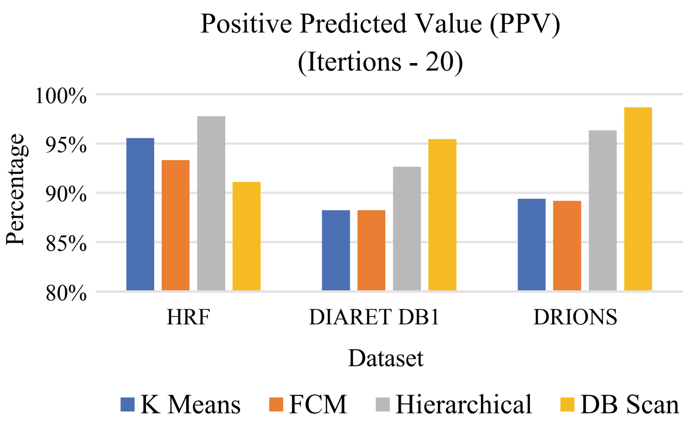


Fig. 29 Positive predicted values of clustering techniques on DRIONS, DIARETDB1 and HRF dataset for 20 iterations

The False Discovery Rate of the clustering algorithms on all the three datasets are determined and is tabulated below. Figures 30 and 31 illustrate that the probability of negatively predicted values is healthier in hierarchical clustering on HRF dataset and DBSCAN clustering for DIARETDB1 and DRIONS dataset by considering it for both 10 and 20 iterations respectively.

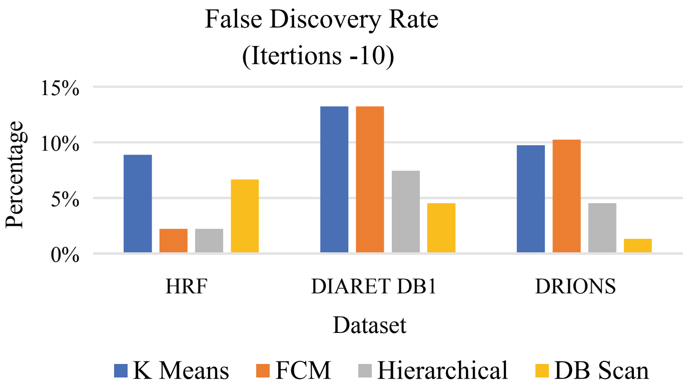


Fig. 30 False discovery rate of clustering techniques on DRIONS, DIARETDB1 and HRF dataset for 10 iterations

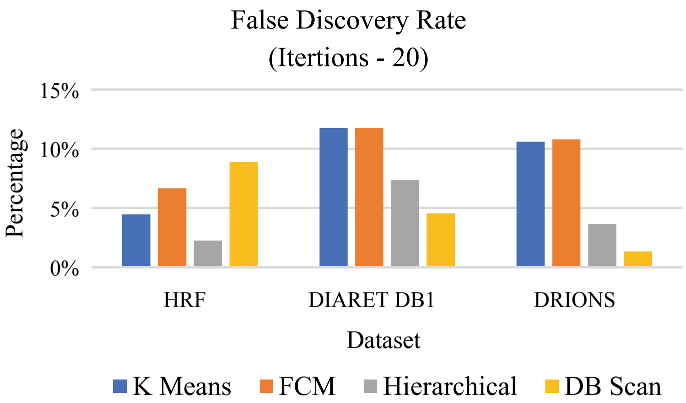


Fig. 31 False discovery rate of clustering techniques on DRIONS, DIARETDB1 and HRF dataset for 20 iterations

6 Conclusion

In this chapter the localization of optic disc is performed on the fundus images obtained from three dataset namely High-Resolution Fundus, DIARETDB1 and DRIONSDB for the iterations 10 and iteration 20 using the unsupervised learning clustering methods like DB Scan clustering, Hierarchical clustering, Fuzzy C Means clustering and K-Means clustering. The results implied that the accuracy of clustering techniques is higher on HRF dataset. Also, it is determined that Precision, Recall, F1 Score of clustering techniques on HRF dataset is better than DIARETDB1 dataset and DRIONSDB dataset for both 10 iterations and 20 iteration. The Positive Predicted Value and False Discovery Rate obtained are also found that the performance of HRF dataset is better than other two datasets. The experimentation results inferred that the clustering techniques on HRF dataset is better when compared to DIARETDB1 and DRIONS dataset. In the future work, the localization of optic disc can also be carried out in a much more efficient way with the help of metaheuristic optimization techniques like Firefly Algorithm, Ant Colony optimization, Genetic algorithms, Differential evolution, etc.,

References

1. Jonas, J.B., Gusek, G.C., Naumann, G.O.: Optic disc, cup and neuro retinal rim size, configuration and correlations in normal eyes. Invest. Ophthalmol. Vis. Sci. **29**(7), 1151–1158 (1988)
2. Marrugo, Andrés, Millán, María: Optic disc segmentation in retinal images. Optica Pura Aplicada **43**, 79–86 (2010)
3. Kusuma Whardana, A., Suciati, N.: A simple method for optic disk segmentation from retinal fundus image. Int. J. Image Graph. Signal Process. **6**(11), 36–42 (2014). https://doi.org/10.5815/ijigsp.2014.11.05
4. Vinoth Kumar, B., Karpagam, G., Zhao, Y.: Evolutionary algorithm with Memetic search capability for optic disc localization in retinal fundus images. Intell. Data Anal. Biomed. Appl. 191–207 (2019). https://doi.org/10.1016/b978-0-12-815553-0.00009-4
5. Wang, L., Liu, H., Zhang, J., Chen, H., Pu, J.: Automated segmentation of the optic disc using the deep learning. Med. Imaging: Image Process. (2019). https://doi.org/10.1117/12.2510372
6. Edupuganti, V.G., Chawla, A., Kale, A.: Automatic optic disk and cup segmentation of fundus images using deep learning. In: 2018 25th IEEE International Conference on Image Processing (ICIP) (2018). https://doi.org/10.1109/icip.2018.8451753
7. Sharma, N., Verma, D.A.: Segmentation and Detection of Optic Disc Using Kmeans Clustering (2015)
8. Khalid, N.E., Noor, N.M., Ariff, N.M.: Fuzzy C-means (FCM) for optic cup and disc segmentation with morphological operation. Procedia Comput. Sci. **42**, 255–262 (2014). https://doi.org/10.1016/j.procs.2014.11.060
9. Lupascu, C.A., Tegolo, D., Rosa, L.D.: Automated detection of optic disc location in retinal images. In: 2008 21st IEEE International Symposium on Computer-Based Medical Systems (2008). https://doi.org/10.1109/cbms.2008.15
10. Youssif, A.A., Ghalwash, A.Z., Ghoneim, A.A.: Optic disc detection from normalized digital fundus images by means of a vessels' direction matched filter. IEEE Trans. Med. Imaging **27**(1), 11–18 (2008). https://doi.org/10.1109/tmi.2007.900326
11. Zhu, X., Rangayyan, R.M.: Detection of the optic disc in images of the retina using the Hough transform. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2008). https://doi.org/10.1109/iembs.2008.4649971
12. Lu, S.: Accurate and efficient optic disc detection and segmentation by a circular transformation. IEEE Trans. Med. Imaging **30**(12), 2126–2133 (2011). https://doi.org/10.1109/tmi.2011.2164261
13. Welfer, D., Scharcanski, J., Kitamura, C.M., Dal Pizzol, M.M., Ludwig, L.W., Marinho, D.R.: Segmentation of the optic disk in color eye fundus images using an adaptive morphological approach. Comput. Biol. Med. **40**(2), 124–137 (2010). https://doi.org/10.1016/j.compbiomed.2009.11.009
14. Yin, F., Liu, J., Ong, S.H., Sun, Y., Wong, D.W., Tan, N.M., Cheung, C., Baskaran, M., Aung, T., Wong, T.Y.: Model-based optic nerve head segmentation on retinal fundus images. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2011). https://doi.org/10.1109/iembs.2011.6090724

15. Cheng, J., Liu, J., Wong, D.W., Yin, F., Cheung, C., Baskaran, M., Aung, T., Wong, T.Y.: Automatic optic disc segmentation with peripapillary atrophy elimination. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2011). https://doi.org/10.1109/iembs.2011.6091537
16. Tjandrasa, H., Wijayanti, A., Suciati, N.: Optic nerve head segmentation using Hough transform and active contours. TELKOMNIKA (Telecommun. Comput. Electron. Control), **10**(3), 531 (2012). https://doi.org/10.12928/telkomnika.v10i3.833
17. Fraga, A., Barreira, N., Ortega, M., Penedo, M.G., Carreira, M.J.: Precise segmentation of the optic disc in retinal fundus images. In: Computer Aided Systems Theory—Eurocast 2011: 13th International Conference, Las Palmas de Gran Canaria, Spain, February 6–11, 2011, revised selected papers, pp. 584–591. Springer (2012)
18. Zhang, D., Yi, Y., Shang, X., Peng, Y.: Optic disc localization by projection with vessel distribution and appearance characteristics. In: 21st International Conference on Pattern Recognition (ICPR 2012) November 11–15, 2012. Tsukuba, Japan, pp. 3176–3179 (2012)
19. Sathya, B., Manavalan, R.: Image segmentation by clustering methods: performance analysis. Int. J. Comput. Appl. **29**(11), 27–32 (2011). https://doi.org/10.5120/3688-5127
20. Ayub, J., Ahmad, J., Muhammad, J., Aziz, L., Ayub, S., Akram, U., Basit, I.: Glaucoma detection through optic disc and cup segmentation using K-mean clustering. In: 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube) (2016). https://doi.org/10.1109/icecube.2016.7495212
21. Wang, W., Zhang, Y., Li, Y., Zhang, X.: The global fuzzy C-means clustering algorithm. In: 2006 6th World Congress on Intelligent Control and Automation (2006). https://doi.org/10.1109/wcica.2006.1713041
22. Hamednejad, G., Pourghassem, H.: Retinal optic disk segmentation and analysis in fundus images using DBSCAN clustering algorithm. In: 2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME) (2016). https://doi.org/10.1109/icbme.2016.7890942
23. Prakash, J.: Enhanced mass vehicle surveillance system. J. Res. **4**(3), 5–9 (2018)
24. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust vessel segmentation in fundus images. Int. J. Biomed. Imaging **2013**, 1–11 (2013). https://doi.org/10.1155/2013/154860
25. Kauppi, T., Kalesnykiene, V., Kamarainen, J., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Uusitalo, H., Kalviainen, H., Pietila, J.: The DIARETDB1 diabetic retinopathy database and evaluation protocol. Proc. Br. Mach. Vis. Conf. (2007). https://doi.org/10.5244/c.21.15
26. Carmona, E.J., Rincon, M., Garcia-Feijoo, J., Martinez-de-la-Casa, J.M.: Identification of the optic nerve head with genetic algorithms. Artif. Intell. Med. **43**(3), 243–259 (2008)
27. Almazroa, A., Burman, R., Raahemifar, K., Lakshminarayanan, V.: undefined. J. Ophthalmol. **2015**, 1–28 (2015). https://doi.org/10.1155/2015/180972

Mr. J. Prakash working as an Assistant Professor in the department of Computer Science & Engineering at PSG College of Technology, Coimbatore. He was born at Coimbatore in Tamil Nadu. He joined BE Compute Science & Engineering at Hindustan Institute of Technology, Coimbatore and received his UG degree on BE Computer Science & Engineering from Anna University Chennai in 2012. He also completed his ME Software Engineering at PSG College of Technology and received his on ME Computer Science & Engineering from Anna university Chennai in 2014 and currently pursuing his Ph.D. degree from Anna University, Chennai. He completed his internship at Infosys Limited, Mysore in 2014 and joined as an assistant professor in the department of Computer Science & Engineering at PSG College of Technology, Coimbatore in 2014. He also has organized many workshops, conferences and also have participated in many technical programs organized by various Institutions around India. He has guided 5 UG project and 2 PG projects and also have mentored two batch of UG students from Assam engineering College, Guwahati.

He has published more than 10 papers in peer reviewed National/International Journals, Conferences and Book chapters. He also has organized many workshops, Conferences. His current area research interests include Digital Image processing, Evolutionary Techniques.

Dr. B. Vinoth Kumar was born in Tamil Nadu, India, in 1981. He received the B.E. degree in Electronics and Communication Engineering from the Periyar University, Tamil Nadu in 2003, and the ME and Ph.D. degrees in Computer Science and Engineering from the Anna University, Tamil Nadu, India, in 2009 and 2016, respectively. In 2004, he joined the Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, as a Lecturer, and in 2010 became an Assistant Professor. Since June 2018, he has been with the Department of Information Technology of the same college, as an Associate Professor. His current research interests include Computational Intelligence, Memetic Algorithms and Digital Image Processing. He has established an Artificial Intelligence Laboratory (AIR Lab) at PSG College of Technology. He is a Life Member of the Institution of Engineers-India (IEI), International Association of Engineers (IAENG) and Indian Society of Systems for Science and Engineering (ISSE). He has published papers in peer reviewed National/International Journals and Conferences and a Reviewer of International Journals.

Soft Computing Techniques in Data Science

Performance Evaluation of Hybrid Machine Learning Algorithms for Medical Image Classification

N. T. Renukadevi

Abstract Medical imaging is the process of creating images of parts of human body for diagnosis and treatment purposes. These images are collected from traditional X-ray based methods like Mammography and Computed Tomography (CT). Some advanced sources of images include Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). Since large volumes of digital medical images are deposited in repository, it is a humongous task to access these voluminous images. To access the required image representation from data store, a technique called Content Based Image Retrieval (CBIR) is currently in use. Content-based image retrieval (CBIR) is an assembly that can overcome the problem mentioned above as it is based on the visual analysis of contents that are part of the query image. CBIR retrieves the images which are needed based on its visual contents. CBIR includes Feature extraction and Feature matching. In feature extraction, information like colour, texture and shape known as feature vectors are retrieved through various extraction methods. Similarly, in feature matching the extracted features are compared between normal and abnormal images for classification. The major challenge in CBIR is implementing flexible methodologies to process the different images of different characteristics like colour, shape and pattern. At the same time, applications for retrieving images for proper indexing is done through Picture Archiving and Communication Systems (PACS). In this chapter, retrieving medical images from different data stores and performance of various machine learning classifiers such as Support Vector Machine (SVM) and Deep Learning methodology are focussed to improve the classification accuracy.

Keywords Support vector machine · Deep learning · Particle swarm optimization · Genetic algorithm · Grasshopper optimization

N. T. Renukadevi (✉)
Department of CT-UG, Kongu Engineering College, Erode, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_12

1 Introduction

In images, features like contrast, brightness, entropy may influence the effectiveness of the entire system. Features like colour, texture, and shape orientation do not match with label manual or human interface. The semantics of images can thus be obtained. Semantic gap occurs due to the variations among these features. In order to make connectivity between the local and global features CBIR faces lot of challenges. Amongst which identifying the required storage area for accumulating the images is the greatest task of CBIR. Next, the performance of the system may also degrade due to high cost incurred in computation of the images.

Medical databases contain digital medical images which includes information about patient's record and image description. These descriptions are utilized for both research and academic processes. Since large volume of digital medical images are created and accumulated in repository, retrieval of particular image is a major risk. Hence, CBIR aids in such retrieval process based on user's query. Previously several researches have been done in CBIR with excellent feature extraction and classification methods like [6, 22]. Research has been done to a large extent to manage particular type of images like mammography, brain tumour [3], lung cancer detection [28] or some blood related diseases [39].

Global features play a vital role in identifying the required images for retrieval [32] when compared to low level features. In this chapter, retrieving medical images from different data bases is focussed.

The key idea of the investigation is to find out the efficiency of machine learning algorithms like Support Vector Machine (SVM) hybridized with optimization techniques such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Deep Learning Techniques.

Further sections focus on:

- Evaluation of the performance of SVM with Radial Basis Function (RBF) Kernels for classifying medical images with Coiflet wavelets and Moment Invariant (MI).
- Implementation of Principal Component Analysis (PCA) for reducing the features of an image.
- Hybrid implementation of optimization algorithms GA and PSO with SVM and evaluating the classification accuracy for multi class medical images.
- Implementation of Deep Learning Technique with Grasshopper Optimization (GOA) algorithm for identification of cancer in liver images.

2 Methodologies

2.1 Efficiency of SVM-RBF Kernels for Medical Image Classification

The medicinal information system in current era shows tremendous changes in making verdict of health conditions about a patient with the aid of images from various scanning devices. In these circumstances, CBIR is very helpful for the retrieval of required data based on the query given by the physicians or medical practitioners. This leads to propose a research work with appropriate machine learning technique SVM to classify medical images obtained through CT scans.

Coiflet wavelets, MI methods, PCA and Kernel PCA feature reduction method along with SVM are needed for classifying the images. Even if the images underwent transformations like Rotation, scaling and Translation, they do not change the performance of wavelets and MI methods for feature extraction. But it is really a tedious task in spatial domain because of the natural characteristics of medical images. But wavelet transformation methods are able to accomplish this inadequacy by removing pixel coefficients of high frequency and observing the features of low frequency coefficients at various resolution levels. In order to extract the shape vectors of an image, MI is used since the image differs in shapes.

Coiflet wavelet is used to extract the energy coefficients because of the orthogonal property displayed by this wavelet of X number of coefficients that leads to a shorter filter with one-third of *X* diminishing instants and one-third of X − 1 scale functions. The wavelet function (Ψ) has twice the *X* instants leads to 0 and the scale function (φ) has $\Psi - 1$ instant leads to 0. Both Ψ and φ hold up a length of six times the $X - 1$ values.

Let *p* represents a standard uninterrupted instant signal, coefficient for maximum value of *i* is $<p, \varphi_{i,k}>$ is computed by $2^{-\frac{i}{2}} p\left(2\varphi_k^i\right)$ Where *p* is a multinomial of degree *n*, the scale function is equal when $n \leq X - 1$. This characteristic is helpful for finding the variation over the function φ with values *i, k* for the sample signal given.

The characteristics of images extracted are organized through SVM for their efficacy [9, 19, 37].

2.1.1 Moment Invariants

MI is applied as an input characteristic in various research areas like processing images, identifying diverse shapes and remote sensing. Moments endow with important description of an entity which unambiguously symbolizes its shape. Identification of constant shapes is done by the categorization in n-dimensional feature space. Conventionally, MI is calculated based on the details inferred by marginal and internal area of the shapes. The moments applied to create invariants are calculated as discrete values practically [9, 19, 35].

Let $f(a, b)$ be a function, then the standard instants are explained by [21] in "(1)":

$$MI_{ab} = \iint a^m b^n f(a, b) \partial a \partial b \tag{1}$$

where MI_{ab} represents the 2D instance of the function $f(a, b)$. The instance sequence is defined as $(m + n)$ and m, n here are natural numbers. This can be shown in "(2)" as distinct values like:

$$MI_{m,n} = \sum_a \sum_b a^m b^n f(a, b) \tag{2}$$

The translation process in an image is normalized with the help of central pixels called as centroid. The pixel positions of an image centroid are computed using the following Eqs. "(3–6)":

$$\bar{a} = \frac{MI_{10}}{MI_{\infty}} \tag{3}$$

$$\bar{b} = \frac{MI_{01}}{MI_{\infty}} \tag{4}$$

The centroid positions are interpreted as:

$$\mu_{mn} = \sum_a \sum_b (a - \bar{a})^m (b - b)^n \tag{5}$$

In order to transform the scaling points, the instances are regularized as

$$\eta = \frac{\mu_{mn}}{\mu'_{\infty}} \tag{6}$$

The normalized centroids are calculated as a group of seven moment invariants which are explained as in "(7)" and are independent of rotation transformation [23].

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} + \eta_{02})^2 + 4\eta^2 11 \\
\phi_3 &= (\eta_{30} + \eta_{12})^2 + (\eta_{03} - \eta_{21})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{03} + \eta_{21})^2 \\
\phi_5 &= (3\eta_{30} - \eta_{12})(\eta_{30} + \eta_{12})^2[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + 3(\eta_{21} + \eta_{03})(\eta_{21} + \eta_{03}) \times [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{20})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + 3(\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \times [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{7}$$

2.1.2 Support Vector Machines

The classification technique is employed in several domains like image analysis, text categorization, and bio-informatics. In order to design a good model using SVM, the challenge is the selection of proper features. Limited number of input features in a classifier produces good computational model. The essential task here is to identify right values of parameters along with proper selection of features, which helps in enhancing the performance of SVM in classification. For classification, kernel function should be chosen and kernel parameters and soft margin constant C must be determined. In this work, the kernel function which is used is RBF Kernel function. Thus, the cost parameter and the kernel parameter should be optimized [24, 26].

Let (p_i, q_i) be a training set with $i, j \in 1, 2 \ldots l$ *where* $p_i \in R^n$ and $q_i \in (1, -1)$. The optimization problem is solved by using SVM as given in "(8)" [17]

$$\min\left(\frac{1}{2}\right)we^T we + c\sum_{i=1}^{l}\varepsilon_i$$
$$\text{subject to } y_i(we^T\varphi(x_i) + b) \geq 1 - \varepsilon_i \text{ and } \varepsilon_i \geq 0 \tag{8}$$

where x_i vector is mapped in elevated dimensional space with the function φ. The margin constant c and the positive slack variable ε_i are crucial in reducing the training errors. The value of c should be a positive whole number. To compute the efficient model for SVM with linear separable hyper plane, Lagrangian method is used by maximizing the objective function which is shown in "(9)":

$$Max_{\infty} = L_D(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j < x_i x_j >$$
$$\text{subject to}: \ 0 \leq \alpha_i \leq C \ i = 1, \ldots, m \text{ and } \sum_{i=1}^{m}\alpha_i y_i = 0 \tag{9}$$

where the constants α_i are called Lagrange multipliers.

The best possible hyper plane is found by maximizing the objective function α_i subject to the constraints $\sum_{i=1}^{m}\alpha_i y_i = 0$ and value of α_i lies between 0 and c. The user should examine the maximum limit on α_i. The product of ordered pair $\alpha(x_i x_j)$ for input data is computed using the kernel function K which is given as "(10)".

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|)^2 \ \textit{where}\ \gamma > 0 \tag{10}$$

2.2 Feature Reduction using PCA

The most acceptable linear method for reducing the dimensionality of image features is PCA. It increases the variance and reduces the data from high dimension feature to lower one. It reduces all the aspect of features with the support of vector space which works well for even complex data sets. In real datasets, empirical zero mean assumed in PCA may not be possible [5]. To overcome this issue, a modification of the PCA kernel is proposed. The data $x_1, x_2, \ldots x_m \in X$ is given and is assumed as a vector area. The principal components $\varphi(x_i), \ldots, \varphi(x_n)$ have been figured out by using Kernel PCA. In general, a problem in PCA should be altered with the requisites of kernel.

2.2.1 Kernel PCA

For a non-empty set N and a non-negative specific kernel k, a function ($N \times$ N) ϵ R has a characteristic that a function $\varphi : n \to H$ is mapped for all $n, n' \in N$ with $(n, n') = k(n, n')$. K can be viewed as a non-linear similarity measure for the kernel methods [12].

At the end, covariance matrix of H will be written as follows in "(11–15)" even in the case of unrestricted dimensions also,

$$CM : \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)^T \tag{11}$$

where $\varphi(x_i)^T$ is the sequential representation of mapping a vector v to the function

$$\sum_{i=1}^{n} \varphi(x_i)^T \text{ and } Cv = \tau v \tag{12}$$

with $\tau \neq 0$ reclined in the distance of φ training images. Hence

$$v = \sum_{i}^{n} \alpha_i \varphi(x_i)^T \tag{13}$$

and this reduces the problem of finding α_i. And α_i is represented in the way of

$$n\tau\alpha = K\alpha_i \tag{14}$$

where $\alpha = (\alpha_i^T)$ and $K = k(x_p, x_q)$. Taking into consideration of these factors in eigen value τN, the zth extractor of the feature will be

$$(v^z, \varphi(x)) = \frac{1}{\sqrt{\tau^z}} \sum_{i=1}^{n} \alpha_i^z k(x_p, x_q) \tag{15}$$

And formulated by finding the multiplication of sample data point $\varphi(x)$ and the zth eigen vector in search space, then $\frac{1}{\sqrt{\tau^z}}$ makes sure that the ordered pair $(v^z, v^z) = 1$.

Thus, KPCA derives zth feature values which is comparatively equal to the coefficient values α_n^z.

Thus, from the Eqs. "(11–15)", a conclusion is drawn that there will be zero mean values for a data in feature space which is shown in "(16)"

$$M_n = \frac{1}{n} \sum_i \varphi(x_i) \tag{16}$$

The difference between mean value Mn and all data points are calculated, which leads to a dissimilar eigen value and its diagonal values can be used as in "(17)"

$$K = (I - ee^T)K(I - ee^T)\text{with } e \text{ as } \frac{1}{n}(T) \text{ rather than } k. \tag{17}$$

2.3 Hybrid Algorithm Based on PSO, GA with Local Search

Among the different amalgamation of parameters, the suitable parameters are selected and retrieved through computations which will be applied for the needed dataset. In order to automate this process, several research work have been carried over on retrieval and optimization techniques for images [20]. And, in continuation to the review, this chapter focuses on hybridizing PSO and GA with various combinations of parameters C and γ, in SVM RBF Kernel function.

GA is an optimization method involved in finding the best possible solutions of search problems. It is a nature inspired evolutionary algorithm to find the optimized solution using heuristics. It produces the probable results for optimization problems by working with small genetic chromosome like data configuration. GA includes stages like selection, crossover, mutation which are applied over these data structures to maintain crucial information [17].

PSO works is based on the communal behaviour of bees or bird congregates/fish groups. It is uncomplicated to implement and produces the results easily. Here, every unique element i denotes mixture of parameters representing the location of element i in exploration area. The velocity of a particle or element shows the route of searching the food items and keep informed the location and velocity of element at regular intervals of iteration. This leads in locating the most excellent area in search space [20]. The location and speed of a particle are computed as given in the Eqs. "(18, 19)"

$$s_i^{di+1} = w.s_i^{di} + ct_1.rnd_1(p_i^{di} - x_i^{di}) + ct_2.rnd_2(p_g^{di} - x_i^{di}) \tag{18}$$

$$x_i^{di+1} = x_i^{di} + s_i^{di} \tag{19}$$

where w is mass weight; di characterizes the number of iterations; i represents the extent of inhabitants; the two "most excellent" values—p_i^{di} means the finest way out accomplished by particle up to the current iteration and p_g^{di} means the finest way out accomplished by any of the particles in the group of inhabitants and will be shared among them. The parameters ct_1 and ct_2 are non-negative constants which evaluate the significance of comprehensive learning of the flock. Also, rnd_1 and rnd_2 are arbitrary values with a range of 0 and 1.

The evaluation of SVM along with RBF kernel depends on two constraint values such as C *and* γ. The values of C and γ are inversely proportional to each other since the increase in value of γ leads to better accuracy and at the same time produce unfair result. Similarly, the increase of C leads to poor accuracy but produce low bias result. Thus, SVM is influenced by these two parameters [4, 17]. In order to make a model of SVM to fit with the training data, optimized values of C *and* γ should be utilized. Such optimization can be done with the support of PSO which looks for the best combination of those parameters with the intention to reduce Root Mean Squared Error (RMSE).

In PSO, the efficiency of optimization is influenced by the parameters w, ct_1, ct_2, rnd_1 and rnd_2 [1, 7, 14]. The weight parameter w manages the searching process and it lies between 0 and 1 so that the elements can congregate. While w increases nearer to 1, it may lead to global exploration while it decreases to 0 from 0.5 it leads to local exploration. The parameters rnd_1 and rnd_2 lies in the range from 0 to 1. The values of ct_1, ct_2 are almost equivalent to each other and they lay in between 0 and 4. The Fig. 1 explores the PSO optimization with SVM.

Rarely these parameter values may have sudden convergence speed, and in order to prevent from such impulsive convergence the PSO algorithm is customized with GA. Since GA already has several phases like selection, crossover and mutation, these phases coordinate with constraints in PSO to improve the performance and to reduce the immature convergence of the particles. In GA, the individual inputs are encoded using binary encoding with two points cross over rate and random mutation with Roulette wheel selection. As GA reaches the end, those parameter coefficients are modified and given as an input to PSO algorithm.

In this moment, GA and PSO are hybridized to become an evolutionary technique as it prevents the impulsive convergence of elements in population. GA does extremely well in routing problem and PSO is excellent in fuzzy control systems and neural networks due to its grouping characteristics. Here, both C and γ are optimized to develop inhabitants based on these optimization algorithms. Also, these exhibits effective appearance while evaluation. During each step, the particles are split into two groups and evaluated with GA and PSO individually. Later, all the particles are pooled into a new population and the previous step is done for next iteration. The iteration continues until the best solution is reached [2, 21, 41].

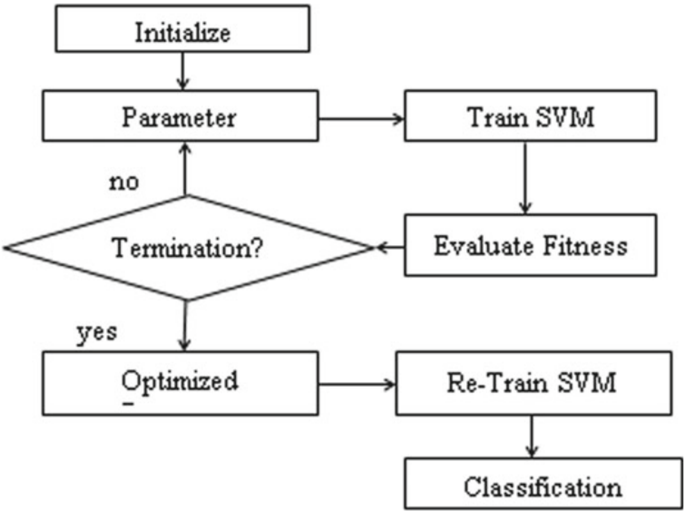


Fig. 1 Flowchart—PSO optimizing SVM

Local Search is the foundation of several heuristics' techniques used to solve hybridized optimization problems. It is a method with frequent repetition of steps that helps to identify solution with good estimation of parameter coefficients. It includes the search space with the parameters as nodes of a graph and objective function as edges. The local search scans the search area by traversing the nodes down the edges.

The important factor to be considered here is the conditions followed by local search algorithm. The conditions are (i) a solitary sequence of data items apart from group-based particles; and (ii) explore the enhanced results in minimized search area defined by self-learning approaches. In general, local search is a repeatedly applied heuristic method whichever optimization algorithm can utilize it [30].

2.4 Deep Learning

Even though SVM is good in classification of medical images both linearly and non-linearly, finding appropriate kernel for complex datasets consumes long time for training. In order to overcome these conditions, Deep Learning technique, which is a compartment of Machine Learning, is being used. It is an inspiration of Neural Networks which is similar to the functioning of human brain and works well on prediction.

A neural network possesses layers like input layer, an intermediary hidden layer, and the last output layer, where each layer consists of "nodes". The medical images collected from different modalities like CT, Mammogram, MRI, PET and Ultrasound are converted to pixels or picture elements which are keyed into the nodes in input layer. Some weights along with input values are fed into hidden layer for computation, and finally the prediction outcome can be obtained from output layer.

Deep learning automatically learns to generate and reduce the pixel characteristics of images, and, if needed, create new features. In addition to this, deep learning method predicts the occurrences of diseases with statistical measures and benefits the practitioners to make decision at the earliest [24, 27]. This chapter also includes an analysis of wide-ranging applications in the field of medical image classification.

2.4.1 Applications of Deep Learning in Medical Image Analysis

Massive computational power and challenges in the investigation of medical imaging is accomplished all the way through deep learning. In the recent years, image detection, segmentation and disease classification are highly involved with deep learning. CNN plays a dynamic role along with selected features among the different deep learning methods.

Clinical practice is improved through deep learning and the illustration is increasing on a daily basis. Deep learning is applied in treatment of disease in the form of radiation [31]. Scanning of images are done through PET or MRI to get detailed pictures of anatomy of body [11, 25, 29], in methods that extracts large volume of features from radiography [13, 34], and in the field which combines diagnostic test with therapy in neurosurgical imaging [10, 15].

When deep learning is applied in healthcare industries, it provides preferable solutions to a variety of problems like diagnosis of diseases, suggestions for personalized treatments etc. A good amount of data is generated through various methods of radiological imaging. But still there are undersized amount of significant data which are needed to be included by means of deep learning model.

2.4.2 Deep Learning in Liver Disease Identification

Nowadays, the frequency and death rate of liver disease are greater than ever. Liver is infected by many types of diseases like Cirrhosis, the scar in liver; Hepatitis, the inflammation in cells; Cholestasis, the obstruction of bile flows, etc., Identifying the disease through CT images is done effectively with the help of deep learning algorithm. In this work, Deep Learning Neural Network DBN incorporated with Grasshopper optimization algorithm is applied for the classification of livers from diseased one [38].

Initially, pre-processing work is done with feature extraction and feature reduction process. For extraction of texture features, Gabor filter along with Local binary Pattern (LBP) is used. Gabor filter detects edge based on frequency values in the image around the point of analysis. LBP works by classifying the pixels based on threshold value of neighbourhood pixels and the output is a binary value. Colour features can be extracted using measures like mean, variance, skewness and kurtosis.

Feature reduction process takes place using PCA which is explained above [36]. Gray Level Co-occurrence Matrix (GLCM) feature extraction method represents the number of incidences of grey intensity values in an image. Features like contrast, entropy, homogeneity, etc., were extracted [36, 38, 42]. Mean value of each region is calculated by using intensity value as expressed in Eq. "(20)"

$$Mean = \frac{1}{NXP} \sum_{A=1}^{N} \sum_{B=1}^{P} f(A, B) \tag{20}$$

Contrast is the variation of colours in an object that makes it distinguishable from others. Equation "(21)" is of the form

$$Con = \sum_{A=1}^{n} \sum_{B=1}^{n} (A - B)^2 C_{AB} \tag{21}$$

where *A, B* represents the luminance and intensity of an object and C represents the Co-occurrence matrix values. If the value of difference of *A* from B is 0 then there will be no contrast, if the difference increases, then the contrast also increases exponentially.

Entropy is the random distribution of intensity values, and when the same values are repeated for certain patterns, entropy is said to be uniform as mentioned in Eq. "(22)". If the entropy is low, then randomness is also low.

$$Entr = \sum_{i}^{n} \sum_{j}^{n} p_{ij} \log p_{ij} \tag{22}$$

where p_i is the intensity value of pixel in a grey scale image.

Homogeneity shows nearness of the pixels of an image as represented in Eq. "(23)"

$$Homo = \sum_{p,q=1}^{n} H_{pq} \left(\frac{1}{1 + (p - q)^2} \right) \tag{23}$$

After the feature extraction and reduction processes, the liver images undergo classification process using Grasshopper Algorithm based Deep Belief Network (DBN). It is a new model generated from Restricted Boltzmann Machines (RBM). Here, the nodes in each layer are connected to previous and subsequent layers, and are used to extract the features of images and classify it [18, 40].

DBN works by taking probability values as input and produces outcome by implementing unsupervised learning algorithm. DBN consists of nodes in input layer, hidden layer and output layer. The input layer includes I ε $\{0, 1\}^{a}$ with parameter values m of any real numbers. The hidden layer contains of J ε $\{0, 1\}^{b}$ with parameter values n of any real numbers. And, *I, J* represent the count of input and hidden nodes

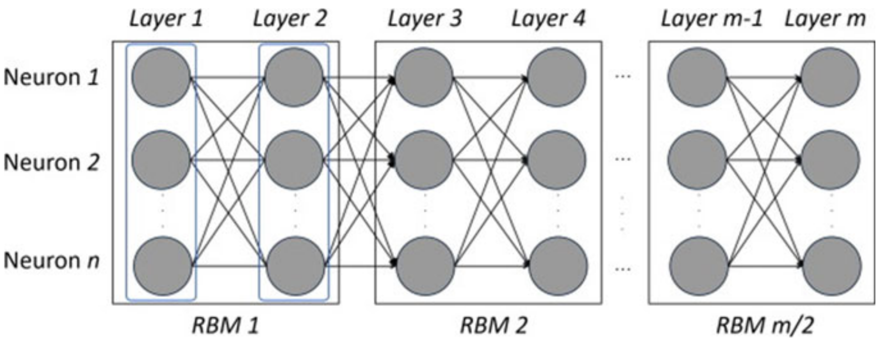


Fig. 2 Architectural diagram of DBN

with weight *Wab*. The main point to note here is, there is no relationship between nodes in a layer on its own. The energy can be obtained by training the parameters *m, n, W*. The formula for computing energy is given below as Eq. "(24)". The architectural diagram of DBN is also depicted in Fig. 2 [8, 18].

$$E(I, J) = -\sum_{a}^{x} m_a I_a - \sum_{b}^{x} n_b J_b - \sum_{i}^{x} \sum_{j}^{x} I_a W_{ab} J_b \tag{24}$$

At first, the initial weights are identified by unsupervised pre-training process in DBN which shows error detection and good optimization results. The best possible values of count of layers and nodes are selected based on the data set type and still there is a deficient in global optimum value. This research work overcomes this problem by implementing Grasshopper Optimization Algorithm (GOA).

GOA is an innovative and efficient optimization algorithm that reduces the movement of grasshoppers to solve optimization problem. Generally, the insects fly as a group and each one has a meticulous space to others and their progress is calculated as in Eq. "(25)"

$$X_i = C_i + F_i + AS_i \tag{25}$$

where $X_{\rm i}$ is the position of ith grasshopper and C_i the shared communication, F the force of attraction, and AS_i the air stream advection on the ith grasshopper respectively [16]. The random behaviour of the equation can be represented in Eq. "(26)" and $\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_3$ are the random numbers with weight between 0 and 1 [33]

$$X_i = C_i r_i + F_i r_2 + AS_i r_3 \tag{26}$$

GOA includes the following steps in Algorithm 1 [16, 38]:

1. Initializing the random population of RBM parameters with minimum, maximum values and no. of iterations.
2. For each iteration:
   a. Compute the fitness solution of each grasshopper g_i
   b. Renew the location of existing search particle.
   c. Standardize the distance between grasshoppers
   d. Update the position of grasshopper g_i with respect to others
   e. If g_i *is* beyond the limit, correct its position
3. If the solution reaches, update the area of attraction.

Algorithm 1. GOA with DBN

GOA does better than PSO and Genetic Algorithm in speed of convergence and can solve complex optimization techniques with less time for good optimum solution [33].

3 Discussion of Experimental Results

Experiments were carried out with 750 images of 3 classes of CT scan medical images like colon, brain and liver and feature extraction was through use of Coiflet wavelet and MI. The images were acquired from National Biomedical Imaging Archive (https://imaging.nci.nih.gov/ncia). Experiments were conducted for 10-fold cross validations, SVM-RBF algorithm with different C and gamma parameters.

Features from Coiflet and MI were combined with the anticipated artifact imperative fusion procedure after obtaining the Median Absolute Deviation (MAD) among the two features.

Let Ci = $\{ci, 1, ci, 2, ci, n\}$ be the Coiflet coefficient.

Let Mi = $\{mi, 1, mi, 2,, mi, n\}$ be the MI coefficients.

The feature vectors are fused by normalizing the feature vector to obtain C*i* and M*i* using Median Absolute Deviation and taking the average of the same.

The formulae for classification accuracy, specificity, sensitivity and f-measure are given as follows in Eqs. "(27–30)":

$$CAcc = \frac{TPos + TNeg}{TPos + FPos + FNeg + TNeg} \tag{27}$$

$$Spec = \frac{TPos}{TPos + FNeg} \tag{28}$$

Table 1 Classification, accuracy of SVM with Coiflet and MI

Techniques used	Accuracy (%)		
	Coiflet	MI	Coiflet and MI
SVM-RBF with Cost C = 1, $\gamma = 0.1$	83.40	78.791	78.75
SVM-RBF with Cost C = 1, $\gamma = 0.05$	81.85	76.46	81.75
SVM-RBF with Cost C = 0.5, $\gamma = 0.5$	81.86	82.84	82.39

$$Sen = \frac{TPos}{TPos + FPos} \tag{29}$$

$$fscr = \frac{2 * Spec * Sen}{Sen + Spec} \tag{30}$$

where

True Negative (*TNeg*)—Count of relevant estimation that an occurrence is untrue
False Positive (*FPos*)—Count of irrelevant estimations that an occurrence is well-founded
False Negative (*FNeg*)—Number of irrelevant estimations that an occurrence is untrue
True Positive (*TPos*)—Number of relevant estimations that an occurrence is well-founded.

Tables 1 and 2 shows the feature fusion of Coiflet with MI for the SVM classifier produces different accuracies for various cost and gamma functions. Since Cost and gamma functions were selected randomly, the performance of the SVM classifier is not optimal. Hence, work needs to be done in the direction of feature reduction and SVM kernel optimization.

Tables 3 and 4 shows the performance of the classifier as it improved significantly due to feature selection as seen from the experimental results. Comparing these results with previous results when no feature reduction techniques were used, it is observed that the Kernel PCA improves the classification accuracy of the classifiers in the range of 2.00% as given in Table 5. Since there is considerable difference in the classification accuracy of the SVM classifier for various cost and gamma parameter, further investigation needs to be carried out to identify the ideal parameters for the RBF kernel.

From the above results, it is shown that Local Search along with SVM-GA-PSO gives better results when compared to others. Also, Kernel PCA gives classification accuracy of about 1.93% higher compared to PCA of the same method. Similarly, DBN with optimization algorithms like GA, PSO, and GOA are experimented. The classification accuracy calculated for 125 liver images alone and the results are tabulated in Table 6.

Table 2 Precision, Recall and F-Measures of SVM with Coiflet and MI

Methods	Precision			Recall			F-Measures		
	Coiflet	MI	Coiflet and MI	Coiflet	MI	Coiflet and MI	Coiflet	MI	Coiflet and MI
SVM-RBF C = 1, $\gamma = 0.1$	0.834	0.812	0.826	0.8332	0.7891	0.8125	0.8314	0.7688	0.8462
SVM-RBF C = 1, $\gamma = 0.05$	0.8325	0.7652	0.8323	0.8547	0.7653	0.8127	0.8334	0.7856	0.8224
SVM-RBF C = 0.5, $\gamma = 0.5$	0.8405	0.8234	0.8146	0.8236	0.8412	0.8432	0.8165	0.8218	0.8167

Table 3 Classification accuracy of SVM using KPCA

Methods	Classification accuracy (%)	
	PCA	KPCA
SVM-RBF with Cost C = 1, $\gamma = 0.1$	84.57	86.57
SVM-RBF with Cost C = 1, $\gamma = 0.05$	84.83	85.90
SVM-RBF with Cost C = 0.5, $\gamma = 0.5$	82.40	84.56

Table 4 Precision, Recall and F-Measures of SVM using KPCA

Methods	Precision		Recall		F-Measure	
	PCA	Kernel PCA	PCA	Kernel PCA	PCA	Kernel PCA
SVM-RBF Kernel C = 1, $\gamma = 0.1$	0.8652	0.8753	0.8432	0.8552	0.8543	0.8626
SVM-RBF Kernel C = 1, $\gamma = 0.05$	0.8234	0.8256	0.8569	0.8589	0.8385	0.8369
SVM-RBF Kernel C = 0.5, $\gamma = 0.5$	0.8434	0.8465	0.8456	0.8498	0.8435	0.8467

Table 5 Accuracy of SVM with optimization techniques

Methods	Classification accuracy (%)	
	PCA	Kernel PCA
SVM-GA	85.82	87.34
SVM-PSO	86.85	90.18
SVM-GA-PSO	90.56	91.54
SVM GAPSO with local search	92.62	94.53

Table 6 Accuracy of DBN with optimization techniques

Methods	Classification accuracy (%)
DBN-GA	95.25
DBN-PSO	94.36
DBN-GOA	97.38

4 Conclusion

In this chapter, medical images were retrieved through feature extraction techniques like Coiflet and MI. The features are reduced by PCA and Kernel PCA techniques and classified by SVM with RBF kernel functions. Since the parameters *C and Gamma* are randomly chosen, the classification accuracy is little poor in performance.

Hence, in order to optimize the values of *C and Gamma,* optimization algorithms are used such as GA and PSO. Experiments are evaluated with SVM and optimization algorithms individually and also as hybridized one. Results show that SVM with GA and PSO along with local search are also calculated. And, according to the current technological improvement in Deep Learning, DBN is also tried with these optimizers and with Grasshopper Optimization algorithm which shows comparatively better performance in classification.

References

1. Ahmed, M.: A survey on dynamic clustering based colour image segmentation using genetic algorithm. World J. Sci. Technol. **1**(12), 35–41 (2012)
2. Ali, A.F., Tawhid, M.A.: A hybrid particle swarm optimization and genetic algorithm with population partitioning for large scale optimization problems. Ain Shams Eng. J. (Ain Shams University) **8**(2), 191–206 (2017)
3. Arakeri, M.P., Ram Mohana Reddy, G.: An intelligent content-based image retrieval system for clinical decision support in brain tumor diagnosis. Int. J. Multimed. Inf. Retr. **2**(3), 175–188 (2013)
4. Ardjani, F., Sadouni, K., Benyettou, M.: Optimization of SVM multiclass by particle swarm (PSO-SVM. In: 2010 2nd International Workshop on Database Technology and Applications, DBTA2010—Proceedings, (December), pp. 32–38 (2010)
5. Avalhais, L.P., da Silva, S.F., Rodrigues, Jr., J.T.: Image retrieval employing genetic dissimilarity weighting and feature space transformation functions. In: Twenty-Seventh Annual ACM Symposium on Applied Computing, pp. 1012–1017 (2012)
6. Bhende, P., Cheeran, P.A.N.: Content Based Image Retrieval in Medical Imaging, pp. 10–15 (2013)
7. Cho, M.Y., Hoang, T.T.: Feature selection and parameters optimization of SVM using particle swarm optimization for fault classification in power distribution systems. Comput. Intell. Neurosci. (2017)
8. Dai, X., et al.: Deep belief network for feature extraction of urban artificial targets. Math. Probl. Eng. **2020**, 1–13 (2020)
9. Dixit, A., Majumdar, S.: Comparative analysis of coiflet and daubechies wavelet using global threshold for image de-noising. Int. J. Adv. Eng. Technol. **6**(5), 2247–2252 (2013)
10. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, J. Am. Med. Assoc. **316**(22), 2402–2410 (2016)
11. Guo, L., Jiang, Q., Jin, X., Liu, L., Zhou, W., Yao, S., Min, W.: A deep convolutional neural network to improve the prediction of protein secondary structure. Curr. Bioinform. **15**, 1–11 (2020)
12. Ham, J.: A kernel view of the dimensionality reduction of manifolds. In: Twenty-first International Conference on Machine Learning, p. 47 (2004)
13. Han, W., et al.: Deep transfer learning and radiomics feature prediction of survival of patients with high-grade gliomas. Am. J. Neuroradiol. **41**(1), 40–48 (2020)
14. Hitam, N.A., Ismail, A.R., Saeed, F.: An optimized support vector machine (SVM) based on particle swarm optimization (PSO) for cryptocurrency forecasting. Procedia Comput. Sci. (Elsevier B.V.) **163**, 427–433 (2019)
15. Izadyyazdanabadi, M., Belykh, E., Mooney, M.A., Eschbacher, J.M., Nakaji, P., Yang, Y., Preul, M.C.: Prospects for theranostics in neurosurgical imaging: empowering confocal laser endomicroscopy diagnostics via deep learning. Front. Oncol. **8** (2018)

16. Jia, H., et al.: Hybrid grasshopper optimization algorithm and differential evolution for multilevel satellite image segmentation. Remote Sens. **11**(9) (2019)
17. Jrad, N., et al.: Sw-SVM: sensor weighting support vector machines for EEG-based brain-computer interfaces. J. Neural Eng. **8**(5) (2011)
18. Kamada, S., Ichimura, T., Harada, T.: Knowledge extraction of adaptive structural learning of deep belief network for medical examination data. Int. J. Semant. Comput. **13**(1), 67–86 (2019)
19. Kaur, S., Kaur, G., Singh, D.D.: Comparative analysis of haar and coiflet wavelets using discrete wavelet transform in digital image compression. Int. J. Eng. Res. Appl. **3**(3), 1–6 (2013)
20. Kaya, D.: Optimization of SVM parameters with hybrid CS-PSO algorithms for Parkinson's Disease in LabVIEW environment. Parkinson's Disease (2019)
21. Keyes, L., Winstanley, A.: Using moment invariants for classifying shapes on large-scale maps. Comput. Environ. Urban Syst. **25**(1), 119–130 (2001)
22. Kumar, A., et al.: Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. J. Digit. Imaging **26**(6), 1025–1039 (2013)
23. Li, D.: Analysis of moment invariants on image scaling and rotation. In: Innovations in Computing Sciences and Software Engineering, pp. 415–419 (2010)
24. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. (Elsevier B.V.) **42**(Dec 2012), 60–88 (2017)
25. Liu, F., et al.: Technical developments. Deep learning Mr imaging-based attenuation correction for PeT/Mr imaging. Radiology **286**(2) (2018)
26. Lorena, A.C., De Carvalho, A.C.P.L.F.: Evolutionary tuning of SVM parameter values in multiclass problems. Neurocomputing **71**(16–18), 3326–3334 (2008)
27. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift fur Medizinische Physik. Elsevier B.V. **29**(2), 102–127 (2019)
28. Malviya, N., Choudhary, N., Jain, K.: Content Based Medical Image Retrieval and Clustering Based Segmentation to Diagnose Lung Cancer, vol. 10, no. 6, pp. 1577–1594 (2017)
29. Mehranian, A., Arabi, H., Zaidi, H.: Vision 20/20: magnetic resonance imaging-guided attenuation correction in PET/MRI: challenges, solutions, and opportunities. Med. Phys. **43**(3), 1130–1155 (2016)
30. Mesloub, S., Mansour, A.: Hybrid PSO and GA for global maximization. Int. J. Open Probl. Comput. Sci. Math. **2**(4), 597–608 (2009)
31. Meyer, P., et al.: Survey on Deep Learning for Radiotherapy. Computers in Biology and Medicine. Elsevier Ltd. (2018)
32. Nalini, P., Malleswari, B.L.: Local versus global features for medical image retrieval. Int. J. Mod. Trends Eng. Res. **4**(10), 1–6 (2017)
33. Nandal, D., Sangwan, O.P.: Bat algorithm, particle swarm optimization and grasshopper algorithm: a conceptual comparison. Int. J. Res. Anal. Rev. **5**(2), 2116–2120 (2018)
34. Oakden-Rayner, L.: Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework. Sci. Rep. **7**(1) (2017)
35. Renukadevi, N.T., College, K.E.: Performance Analysis of Coiflet Wavelet and Moment Invariant Feature Extraction for CT Image Classification using SVM, vol. 2 no. 12, pp. 1–6 (2013)
36. Renukadevi, N.T., Karunakaran, S., Saraswathi, K.: Gray level cooccurrence matrix feature extraction and fuzzy based discriminative binary descriptor for medical image retrieval. Int. J. Comput. Sci. Eng. **7**(6), 62–70 (2019)
37. Renukadevi, N.T., Thangaraj, P.: Performance Evaluation of SVM-RBF Kernel for Medical Image Classification, vol. 13, no. 4 (2013)
38. Renukadevi, T., Karunakaran, S.: Optimizing deep belief network parameters using grasshopper algorithm for liver disease classification. Int. J. Imaging Syst. Technol. **30**(1), 168–184 (2020)
39. Sanghavi, J., Kayande, D.: Content based image retrieval (CBIR) system for diagnosis of blood related diseases. In: National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013), (February), pp. 11–15 (2013)
40. Tumuluru, P., Ravi, B.: GOA-based DBN: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification. Int. J. Appl. Eng. Res. **12**(24), 14218–14231 (2017)

41. Umamaheswari, J., Radhamani, G.: A hybrid approach for classification of DICOM image. World Comput. Sci. Inf. Technol. J. **1**(8), 364–369 (2011)
42. Varish, N., Pal, A.K.: Appl. Intell. **48**, 2930–2953 (2018)

Dr. N. T. Renukadevi received the M.C.A. and M.Phil. Degrees in Computer Science from Bharathiar University, India in 2003 and 2008 respectively. She received Ph.D. degree in Computer Applications from Anna University, India in 2014. Since 2008, she is working as an Assistant Professor in the Department of Computer Technology, Kongu Engineering College. She has been in the teaching profession for the past 15 years. She has published 17 papers in International Journals and Conferences. Her research interests include Data Mining and Medical Image Processing. She is also a reviewer in reputed journals.

Computing Truth Values of Modus Ponens and Modus Tollens Rule for Linguistic Truth-Valued Propositions and Its Application in Taking Decisions in Health Care

Sumita Basu and Bithi Chattaraj

Abstract Soft computing is used to find solutions of real life problems where exact computation is not possible. Soft computation models are usually approximate in nature. In this chapter we will discuss an approximate reasoning model. The basic reasoning tools in logic are Modus Ponens and Modus Tollens where a statement is either true or false. However our practical experience tells us that often we refer to a statement as *quite true or somewhat false* etc. Reasoning with such statements could not be done with basic tools of logic. We have tried to find the degree of truth or falsity of these reasoning tools for statements having such linguistic truth values. We have assumed that the propositions have linguistic truth values which are not completely ordered but may be represented by a lattice (Fig. 1 or Fig. 2). It was found that Modus Ponens and Modus Tollens rules with such propositions were not tautologies as in classical logic but have linguistic truth values. These truth values have been computed for all cases in this chapter. The results may be applied to physical, biological, economic or social phenomena. We have showed an application of decision making with linguistic truth-valued propositions in healthcare.

Keywords Modus Ponens · Modus Tollens · Lattice implication algebra · Quasi-lattice implication algebra

1 Introduction

Natural language is not crisp but vague. Reasoning with a crisp language is dealt in classical logic where the truth value set(C) is {T, F}. Logicians faced difficulty to draw inference from the sentences of natural language with two-valued logic. In 1920 Lukasiewicz [4] proposed the theory of three valued logic which was later

S. Basu · B. Chattaraj (✉)
Department of Mathematics, Bethune College, University of Calcutta, Kolkata, West Bengal, India

S. Basu
Guest Faculty, Department of Mathematics, The Heritage College, University of Calcutta, Kolkata, West Bengal, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_14

generalized to multi-valued logic. L. A. Zadeh in 1975 introduced linguistic variables [13–15] to capture such vague concepts.

As an example *Age* is a linguistic variable which may have truth values *very young, moderately young, moderately old, very old* etc. Zadeh represented the truth values of the linguistic variable by a fuzzy set and used fuzzy logic for reasoning with linguistic variables. Fuzzy reasoning can be viewed as a fuzzy extension of multi-valued logic.

Construction of suitable fuzzy set for a typical linguistic variable is very difficult. This makes problem of reasoning with linguistic variable (fuzzy reasoning) all the more challenging. Nguyen and Wechler [6–8] tried to give an algebraic structure to the linguistic truth values and applied the results to fuzzy logic. Some modifications of representation of linguistic variables and its application to fuzzy reasoning have been suggested by Di Lascio et al. [2], Nguyen and Huynh [5], Huynh [3], Cock and Kerre [1].

The truth values of propositions of languages in real world are not exactly defined but are tagged with linguistic hedges. So given a proposition P instead of saying that '*the proposition is true*', we very often say P is *absolutely true/highly true/quite true/somewhat true/rather true/slightly true* etc. Similarly, '*the proposition is false*', is replaced by the P is *absolutely false/highly false/quite false/somewhat false/rather false/slightly false* etc. The linguistic hedge set (H) will be {*absolutely, highly, quite, somewhat, rather, slightly*}. So the set truth values (V) of propositions of natural language will be $V = H \times C$.

A mapping of elements of a set A to [0, 1] implies that there is a linear ordering of the elements of A. However, in real world the elements may be incomparable. So, fuzzy set theory is not adequate to deal with such non-comparable information. In fact Zadeh (1965) commented: '*In a more general setting, the range of the membership function can be taken to be a suitable partially ordered set P*'.

A lattice consists of a set of elements which may be comparable or non-comparable.

In this paper we make the following assumptions:

1. The linguistic hedge set (H) is finite and totally ordered.
2. The linguistic truth valued set V forms a lattice having the Hasse diagram given either by Fig. 1 or Fig. 2.

If V forms a lattice of the form Fig. 1 then a unary operation (inverse operation) "$'$" and a binary operation (implication operation) "$\rightarrow$" on V may be defined so that $L = (V, \vee, \wedge, ', \rightarrow, O, I)$ is a lattice implication algebra which have been studied by Xu [10] and Yiquan et al. [12]. However, if V is of the form 14.2 then L is a quasi-lattice implication algebra. A special class of multi-valued logic called lattice-valued logic have been discussed by Xu et al. [10, 11, 16].

In classical logic Modus Ponens affirms that if *P is true* and *P implies Q* is true then *Q is true* while Modus Tollens states if *not Q is true* and *P implies Q is true* then *not P is true.*

Xu et al. defined propositions having linguistic truth values as LTVP or QLTVP and studied the logic with such propositions. We have used such lattice-valued logic

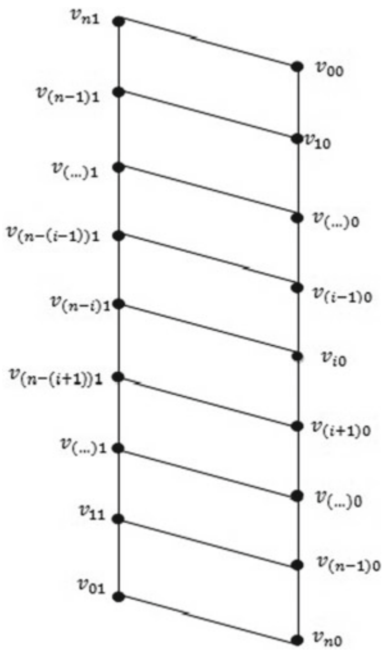


Fig. 1 Hasse diagram of linguistic truth value set V satisfying LIA

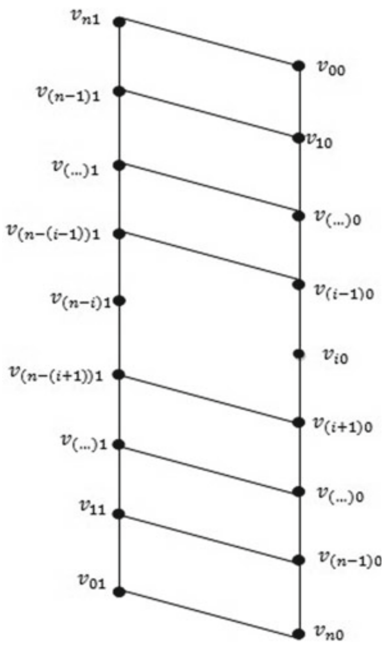


Fig. 2 Hasse diagram of linguistic truth value set V satisfying QLIA

to compute truth values of Modus Ponens and Modus Tollens rules for propositions having the linguistic truth value V. The results show that the truth values of Modus Ponens and Modus Tollens rule are also linguistic truth valued.

The chapter is organized as follows: In Sect. 2 the basic properties of Quasi Lattice Implication algebra (QLIA), Lattice Implication algebra (LIA), Linguistic Truth valued propositions (LTVP), Quasi Linguistic Truth valued propositions (QLTVP) are discussed briefly. Truth values of Modus Ponens and Modus Tollens rules for LTVP and QLTVP are computed in Sect. 3. In Sect. 4 we have reported application of the obtained results to Covid-19 patients. Some concluding remarks are included in Sect. 5.

2 Basic Concepts

The definitions and theorems related to LIA, QLIA, LTVP and QLTVP reported here are taken from [10–12].

Definition 2.1 Let <*L*, ∨, ∧, ′, → , *O*, *I*> be a bounded lattice with universal boundaries *O* (the least element) and *I* (the greatest element) respectively, and "′" be an order-reversing involution. For any $x, y, z \in L$, if mapping $\rightarrow: L \times L \rightarrow L$ satisfies:

1. $(I_1)\ x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z)$
2. $(I_2)\ x \rightarrow x = I$
3. $(I_3)\ x \rightarrow y = y' \rightarrow x'$
4. $(I_4)\ x \rightarrow y = y \rightarrow x = I,\ then\ x = y$
5. $(I_5)\ (x \rightarrow y) \rightarrow y = (y \rightarrow x) \rightarrow x$
   then <*L*, ∨, ∧, ′, → , *O*, *I*> is called a *quasi-lattice implication algebra.*
   If it satisfies two additional properties as follows:
6. $(I_6)\ (x \vee y) \rightarrow z = (x \rightarrow z) \wedge (y \rightarrow z)$
7. $(I_7)\ (x \wedge y) \rightarrow z = (x \rightarrow z) \vee (y \rightarrow z)$
   then <*L*, ∨, ∧, ′, → , *O*, *I*> is called a *lattice implication algebra.*

In classical logic the truth value of a proposition is either true or false. However, in natural language the truth values of statements are not restricted to only true or false; rather they are accompanied by some linguistic hedges which reflect the degrees of truth or falsity of statements. Very often we refer to a statement as somewhat true or slightly false, "somewhat", "slightly" are linguistic hedges.

Let $H = \{h_0, h_1, \ldots, h_n | n \geq 0\}$, be the linguistic hedge set, where $h_0 = slightly,\ h_1 = somewhat,\ h_2 = rather, \ldots,\ h_n = absolutely$. The linguistic hedge operator set *H* is totally ordered.

Theorem 2.1 *Let the hedge operator set* $H = \{h_0, h_1, \ldots, h_n | n \geq 0\}$*, be a chain such that for* $j \leq k,\ h_j \leq h_k;\ h_j \vee h_k = h_{\max(j,k)}$ *and* $h_j \wedge h_k = h_{\min(j,k)}$ *. Now, we define the unary operator "′" as* $h'_j = h_{(n-j)}$ *and the binary operator "→" as* $h_j \rightarrow h_k = h_{\min(n,n-j+k)}$*. Then the set* $\langle H, \wedge, \vee, ', \rightarrow \rangle$ *is a LIA.*

Proof Let $x = h_i,\ y = h_j,\ z = h_k$.

The property I_1 is satisfied as:

$x \rightarrow (y \rightarrow z) = h_{\min(n,n-i+\min(n,n-j+k))}$ and $y \rightarrow (x \rightarrow z) = h_{\min(n,n-j+\min(n,n-i+k))}$

Therefore, $x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z) = \begin{cases} h_n, & if\ j \le k \\ h_n, & if\ j \ge k,\ i \le k \\ h_{(\min(n,2n-i-j+k))}, & if\ j \ge k,\ i > k \end{cases}$

From the definitions of implications given above we can say that the property $I_2:\ x \rightarrow x = h_i \rightarrow h_i = h_n = I$ is satisfied for all $h_i \in H,\ i \in \{0, 1, \ldots, n\}$.

Property I_3, is satisfied as $x \rightarrow y = h_{(\min(n,n-i+j))} = y' \rightarrow x'$.

Property $I_4:\ x \rightarrow y = \ y \rightarrow x = I,\ then\ x = y$, may be easily proved from the definition of implication.

Property I_5 is true in this case as,

$(x \rightarrow y) \rightarrow y = h_{\min(n,n-\min(n,n-i+j)+j)}$ and $(y \rightarrow x) \rightarrow x = h_{\min(n,n-\min(n,n-j+i)+i)}$

Therefore,

$$(x \rightarrow y) \rightarrow y = (y \rightarrow x) \rightarrow x = \begin{cases} h_i,\ if\ i \ge j \\ h_j,\ if\ i < j \end{cases}$$

For the property I_6, since,

$(x \vee y) \rightarrow z = h_{\min(n,n-\max(i,j)+k)}$ and $(x \rightarrow z) \wedge (y \rightarrow z) = h_{\min(\min(n,n-i+k),\min(n,n-j+k))}$, so,

$$(x \vee y) \rightarrow z = (x \rightarrow z) \wedge (y \rightarrow z) = \begin{cases} h_{\min(n,n-i+k)},\ if\ i \ge j \\ h_{\min(n,n-j+k)},\ if\ i < j \end{cases}$$

For the property I_7,

$$(x \wedge y) \rightarrow z = h_{\min(n,n-\min(i,j)+k)}$$

and

$$(x \rightarrow z) \vee (y \rightarrow z) = h_{\max(\min(n,n-i+k),\min(n,n-j+k))}, \qquad \text{(so,)}$$

$$(x \wedge y) \rightarrow z = (x \rightarrow z) \vee (y \rightarrow z) = \begin{cases} h_{\min(n,n-j+k)},\ if\ i \ge j \\ h_{\min(n,n-i+k)},\ if\ i < j \end{cases}$$

Thus, from the above properties we can see that the set $\langle H, \wedge, \vee, ', \rightarrow \rangle$ forms a LIA.

The basic truth value set is *C* = *{T, F}* where *T* = *True* and *F* = *False.*

Let V be the set of all linguistic truth values, i.e. $V = H \times C$. Thus, if v is a linguistic truth value then, $v \in V$ and $v = (h_i, c_j)$ where $h_i \in H,\ c_j \in C$ is composed of a linguistic hedge operator h_i and a basic truth value c_j.

If $V = \{v_{00}, v_{01}, v_{10}, v_{11}, \ldots, v_{n0}, v_{n1}\}$ then $v_{i0} = (h_i, c_0) = (h_i, F)$ and $v_{i1} = (h_i, c_1) = (h_i, T)$. So if h_i represents *"somewhat"* then v_{i0} represents *"somewhat false"* and v_{i1} represents *"somewhat true"*.

Let, $V_1 = \{v_{i1} | i = 0, 1, 2, \ldots, n\}$ and $V_0 = \{v_{i0} | i = 0, 1, 2, \ldots, n\}$ so that $V = V_0 \cup V_1$. Also $V_0,\ V_1$ satisfy the following:

$$v_{i1}, v_{j1} \in V_1,\ i \leq j\ \Rightarrow v_{i1} \leq v_{j1}$$

and,

$$v_{i0}, v_{j0} \in V_0,\ i \leq j\ \Rightarrow v_{i0} \geq v_{j0}$$

So, both V_0 and V_1 are totally ordered and may be represented by chain. Also, the elements of V have the following order

$$k \in \{0, 1, \ldots, n\},\ v_{k0} \leq v_{(n-k)1}$$

Hence V is a partially ordered set and may be represented by the Hasse diagram given by Fig. 1.

Theorem 2.2 *If V be the linguistic truth value set, then* $(V, \leq)$ *is a poset. Let* $\chi = (V, \wedge, \vee, O, I)$ *where* $O = v_{n0},\ I = v_{n1}$ *and* $\forall\ i, j \in \{0, 1, \ldots, n\},\ r = \max\{i, j\},\ s = \min\{i, j\}$. *If the operation "*$\vee$*" and "*$\wedge$*" are defined as*

1. $v_{i1} \vee v_{j1} = v_{r1}$
2. $v_{i0} \vee v_{j0} = v_{s0}$
3. $v_{i1} \vee v_{j0} = \begin{cases} v_{i1}, & if\ n \leq (i + j) \\ v_{(n-j)1}, & if\ n \geq (i + j) \end{cases}$
4. $v_{i1} \wedge v_{j1} = v_{s1}$
5. $v_{i0} \wedge v_{j0} = v_{r0}$
6. $v_{i1} \wedge v_{j0} = \begin{cases} v_{j0}, & if\ n \leq (i + j) \\ v_{(n-i)0}, & if\ n \geq (i + j) \end{cases}$

then χ *forms a lattice.*

Proof In the properties 1, 2, 5 and 6 if we put $i = j$ then we see that the idempotent law is satisfied.

From the above definitions of $\wedge$ and $\vee$ we see that the commutative law is also satisfied.

Now we see whether associative law holds.

(a) Let $x = v_{i1}, y = v_{j1}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$v_{i1} \wedge (v_{j1} \wedge v_{k1}) = v_{i1} = (v_{i1} \wedge v_{j1}) \wedge v_{k1} \; if \; i \le j \le k.$$

and

$$v_{i1} \vee (v_{j1} \vee v_{k1}) = v_{k1} = (v_{i1} \vee v_{j1}) \vee v_{k1} \; if \; i \le j \le k.$$

(b) Let $x = v_{i1}, y = v_{j1}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$v_{i1} \wedge (v_{j1} \wedge v_{k0}) = \begin{cases} v_{k0}, & if \; n \le (i+k) \\ v_{(n-i)0}, & if \; n \ge (i+k) \end{cases} = (v_{i1} \wedge v_{j1}) \wedge v_{k0} \; if \; i \le j \le k$$

and

$$v_{i1} \vee (v_{j1} \vee v_{k0}) = \begin{cases} v_{j1}, & if \; n \le (j+k) \\ v_{(n-k)1}, & if \; n \ge (j+k) \end{cases} = (v_{i1} \vee v_{j1}) \vee v_{k0} \; if \; i \le j \le k$$

(c) Let $x = v_{i1}, y = v_{j0}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$v_{i1} \wedge (v_{j0} \wedge v_{k1}) = \begin{cases} v_{j0}, & if \; n \le (i+j) \\ v_{(n-i)0}, & if \; n \ge (i+j) \end{cases} = (v_{i1} \wedge v_{j0}) \wedge v_{k1} \; if \; i \le j \le k$$

and

$$v_{i1} \vee (v_{j0} \vee v_{k1}) = \begin{cases} v_{k1}, & if \; n \le (j+k); n \ge (i+j) \\ v_{(n-j)1}, & if \; n \ge (j+k); n \ge (i+j) \end{cases}$$
$$= (v_{i1} \vee v_{j0}) \vee v_{k1} \; if \; i \le j \le k$$

(d) Let $x = v_{i1}, y = v_{j0}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$v_{i1} \wedge (v_{j0} \wedge v_{k0}) = \begin{cases} v_{k0}, & if \; n \le (i+j) \\ v_{(n-i)0}, & if \; n \ge (i+k) \end{cases} = (v_{i1} \wedge v_{j0}) \wedge v_{k0} \; if \; i \le j \le k$$

and

$$v_{i1} \vee (v_{j0} \vee v_{k0}) = \begin{cases} v_{i1}, & if \; n \le (i+j) \\ v_{(n-j)1}, & if \; n \ge (i+j) \end{cases} = (v_{i1} \vee v_{j0}) \vee v_{k0} \; if \; i \le j \le k$$

(e) Let $x = v_{i0}, y = v_{j1}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \dots, n\}$

$$v_{i0} \wedge (v_{j1} \wedge v_{k1}) = \begin{cases} v_{i0}, & if\ n \leq (i + j) \\ v_{(n-j)0}, & if\ n \geq (i + j) \end{cases} = (v_{i0} \wedge v_{j1}) \wedge v_{k1}\ if\ i \leq j \leq k$$

and

$$v_{i0} \vee (v_{j1} \vee v_{k1}) = \begin{cases} v_{k1}, & if\ n \leq (i + j) \\ v_{(n-i)1}, & if\ n \geq (i + k) \end{cases} = (v_{i0} \vee v_{j1}) \vee v_{k1}\ if\ i \leq j \leq k$$

(f) Let $x = v_{i0}, y = v_{j1}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \dots, n\}$

$$v_{i0} \wedge (v_{j1} \wedge v_{k0}) = \begin{cases} v_{k0}, & if\ n \geq (i + j);\ n \leq (j + k) \\ v_{(n-j)0}, & if\ n \geq (i + j);\ n \geq (j + k) \end{cases}$$
$$= (v_{i0} \wedge v_{j1}) \wedge v_{k0}\ if\ i \leq j \leq k$$

and

$$v_{i0} \vee (v_{j1} \vee v_{k0}) = \begin{cases} v_{j1}, & if\ n \leq (j + k);\ n \leq (i + j) \\ v_{(n-i)1}, & if\ n \leq (j + k);\ n \geq (i + j) \end{cases}$$
$$= (v_{i0} \vee v_{j1}) \vee v_{k0}\ if\ i \leq j \leq k$$

(g) Let $x = v_{i0}, y = v_{j0}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \dots, n\}$

$$v_{i0} \wedge (v_{j0} \wedge v_{k1}) = \begin{cases} v_{j0}, & if\ \ n \leq (j + k) \\ v_{(n-k)0}, & if\ n \geq (i + k);\ n \geq (j + k) \end{cases}$$
$$= (v_{i0} \wedge v_{j0}) \wedge v_{k1}\ if\ i \leq j \leq k$$

and

$$v_{i0} \vee (v_{j0} \vee v_{k1}) = \begin{cases} v_{k1}, & if\ n \leq (i + k);\ n \leq (j + k) \\ v_{(n-i)1}, & if\ n \leq (j + k);\ n \geq (i + k) \end{cases}$$
$$= (v_{i0} \vee v_{j0}) \vee v_{k1}\ if\ i \leq j \leq k$$

(h) Let $x = v_{i0}, y = v_{j0}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \dots, n\}$

$$v_{i0} \wedge (v_{j0} \wedge v_{k0}) = v_{k0} = (v_{i0} \wedge v_{j0}) \wedge v_{k0}\ if\ i \leq j \leq k.$$

and

$$v_{i0} \vee (v_{j0} \vee v_{k0}) = v_{i0} = (v_{i0} \vee v_{j0}) \vee v_{k0} \ if \ i \leq j \leq k.$$

Therefore, we see that the associative law is also satisfied.

Now, we check for the absorption law.

(a) Let $x = v_{i1}, y = v_{j1}, \forall i, j \in \{0, 1, 2, \dots, n\}$

$$v_{i1} \wedge (v_{i1} \vee v_{j1}) = v_{i1} = v_{i1} \vee (v_{i1} \wedge v_{j1})$$

(b) Let $x = v_{i1}, y = v_{j0}, \forall i, j \in \{0, 1, 2, \dots, n\}$

$$v_{i1} \wedge (v_{i1} \vee v_{j0}) = v_{i1} = v_{i1} \vee (v_{i1} \wedge v_{j0})$$

(c) Let $x = v_{i0}, y = v_{j1}, \forall i, j \in \{0, 1, 2, \dots, n\}$

$$v_{i0} \wedge (v_{i0} \vee v_{j1}) = v_{i0} = v_{i0} \vee (v_{i0} \wedge v_{j1})$$

(d) Let $x = v_{i0}, y = v_{j0}, \forall i, j \in \{0, 1, 2, \dots, n\} v_{i0} \wedge (v_{i0} \vee v_{j0}) = v_{i0} = v_{i0} \vee (v_{i0} \wedge v_{j0})$

Thus, the absorption law is also satisfied. Hence, χ forms a lattice.

Definition 2.2 Let $\chi = (V, \wedge, \vee, O, I)$ be a lattice as defined in the previous theorem. We define $\forall i \in \{0, 1, 2, \dots, n\}, \ j \in \{0, 1\}$ a unary operation (inverse operation) "′" on V as

$$\big((h_i, T)' = (h_i, F)\big); \quad \big((h_i, F)' = (h_i, T)\big)$$

i.e.

$$v'_{ij} = v_{i(1-j)}$$

A binary operation (implication operation) "→" is also defined on V as follows:

1. $\big((h_i, T) \to (h_j, F) = (h_{\max(0,i+j-n)}, F)\big)$ $i.e.\ v_{i1} \to v_{j0} = v_{\max(0,i+j-n)0}$
2. $\Big((h_i, F) \to (h_j, T) = (h_{\min(n,i+j)}, T)\Big)$ $i.e.\ v_{i0} \to v_{j1} = v_{\min(n,i+j)1}$
3. $\Big((h_i, T) \to (h_j, T) = (h_{\min(n,n-i+j)}, T)\Big)$ $i.e.\ v_{i1} \to v_{j1} = v_{\min(n,n-i+j)1}$
4. $\Big((h_i, F) \to (h_j, F) = (h_{\min(n,n-j+i)}, T)\Big)$ $i.e.\ v_{i0} \to v_{j0} = v_{\min(n,n-j+i)1}$

Theorem 2.3 *If* $L = (V, \wedge, \vee, ', O, I, \rightarrow)$ *then L is a lattice implication algebra.*

Proof $(V, \wedge, \vee, ', O, I, \rightarrow)$ forms a bounded lattice due to Theorem 2.2. "′" is an order reversing involution as defined previously.

Property $I_1 : x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z)$ is satisfied as

(a) Let $x = v_{i1}, y = v_{j1}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z) = \begin{cases} v_{\min(n, 2n-i-j+k)1}, & if\ (i, j > k) \\ v_{n1},\ if\ (i, j \le k);\ (j \le k, i > k);\ (j > k, i \le k) \end{cases}$$

(b) Let $x = v_{i1}, y = v_{j1}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z)$$
$$= \begin{cases} v_{\max(0, i+j+k-2n)0}, & if\ (i, j > n-k) \\ v_{00},\ if\ (i, j \le n-k);\ (j \le n-k, i > n-k);\ (j > n-k, i \le n-k) \end{cases}$$

(c) Let $x = v_{i1}, y = v_{j0}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z)$$
$$= \begin{cases} v_{\min(n, n-i+j+k)1}, & if\ (i > k, j \le n-k) \\ v_{n1},\ if\ (i, j \ge n-k);\ (j \ge n-k, i < k);\ (j < n-k, i \le k) \end{cases}$$

(d) Let $x = v_{i1}, y = v_{j0}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z)$$
$$= \begin{cases} v_{\min(n, 2n-i+j-k)1}, & if\ (i > n-k, j < k) \\ v_{n1},\ if\ (i \le n-k, j \ge k);\ (i > n-k, j \ge k);\ (i \le n-k, j < k) \end{cases}$$

(e) Let $x = v_{i0}, y = v_{j1}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z)$$
$$= \begin{cases} v_{\min(n, n+i-j+k)1}, & if\ (i < n-k, j > k) \\ v_{n1},\ if\ (i \ge n-k, j \le k);\ (i < n-k, j \le k);\ (i \ge n-k, j > k) \end{cases}$$

(f) Let $x = v_{i0}, y = v_{j1}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$x \rightarrow (y \rightarrow z) = y \rightarrow (x \rightarrow z)$$
$$= \begin{cases} v_{\min(n, 2n+i-j-k)1}, & if\ (i < k, j > n-k) \\ v_{n1},\ if\ (i \ge k, j \le n-k);\ (i < k, j \le n-k);\ (i \ge k, j > n-k) \end{cases}$$

(g) Let $x = v_{i0}, y = v_{j0}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$\begin{aligned} x \to (y \to z) &= y \to (x \to z) \\ &= \begin{cases} v_{\min(n,i+j+k)1}, & if\ (i, j < n-k) \\ v_{n1},\ if\ (i, j \geq n-k);\ (j \geq n-k, i < n-k);\ (j < n-k, i \geq n-k) \end{cases} \end{aligned}$$

(h) Let $x = v_{i0}, y = v_{j0}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$x \to (y \to z) = y \to (x \to z) = \begin{cases} v_{\min(n,n+i+j-k)1}, & if\ (i, j < k) \\ v_{n1},\ if\ (i, j \geq k);\ (j \geq k, i < k);\ (j < k, i \geq k) \end{cases}$$

Property $I_2 : x \to x = I$ is satisfied as, either $x = v_{i0}\ or\ v_{j1}$.
In either case $x \to x = v_{n1} = I$.
Property I_3 is satisfied as we have the following:

1. $x = v_{i0}, y = v_{j0} \Rightarrow (x \to y = v_{\min(n,n-j+i)1} = y' \to x')$
2. $x = v_{i0}, y = v_{j1} \Rightarrow (x \to y = v_{\min(n,j+i)1} = y' \to x')$
3. $x = v_{i1}, y = v_{j1} \Rightarrow (x \to y = v_{\min(n,n-i+j)1} = y' \to x')$
4. $x = v_{i1}, y = v_{j0} \Rightarrow (x \to y = v_{\max(0,j+i-n)0} = y' \to x')$

Property I_4 is satisfied as we have the following:

1. If $x = v_{i0}, y = v_{j0}$

$$(v_{i0} \to v_{j0}) = (v_{j0} \to v_{i0}) = v_{n1}$$

Now $$\Rightarrow\ v_{\min(n,n-j+i)1} = v_{\min(n,n-i+j)1} = v_{n1}$$

$$\Rightarrow\ (n \leq n - i + j) \wedge (n \leq n - j + i) \Rightarrow i = j$$

2. Similarly, if $x = v_{i1}, y = v_{j1}$

$$\begin{aligned} x \to y = v_{\min(n,n-i+j)1} = y \to x = v_{\min(n,n-j+i)1} &= v_{n1} \\ &\Rightarrow i = j \end{aligned}$$

However, if $x = v_{i1}, y = v_{j0}$ then $x \to y = v_{\max(0,i+j-n)0} \neq v_{n1}$, hence the required conditions are not satisfied.

Satisfaction of I_5 is shown below:

1. Case1: $x = v_{i1}, y = v_{j1}$

$$(x \to y) \to y = v_{\min(n, n - \min(n, n-i+j)+j)1} \ and$$
$$(y \to x) \to x = v_{\min(n, n - \min(n, n-j+i)+i)1}$$

Therefore,

$$(x \to y) \to y = (y \to x) \to x = \begin{cases} v_{i1}, \ if \ j \leq i \\ v_{j1}, \ if \ j \geq i \end{cases}$$

2. Case 2: Similarly, if $x = v_{i0}, y = v_{j0}$

$$(x \to y) \to y = (y \to x) \to x = \begin{cases} v_{j0}, \ if \ j \leq i \\ v_{i0}, \ if \ j \geq i \end{cases}$$

3. Case 3: If, $x = v_{i1}, y = v_{j0}$

$$(x \to y) \to y = (y \to x) \to x = \begin{cases} v_{i1}, \ if \ n \leq i + j \\ v_{(n-j)1}, \ if \ n \geq i + j \end{cases}$$

4. Case 4: If, $x = v_{i0}, y = v_{j1}$

$$(x \to y) \to y = (y \to x) \to x = \begin{cases} v_{j1}, \ if \ n \leq i + j \\ v_{(n-i)1}, \ if \ n \geq i + j \end{cases}$$

Now, let us check the properties $I_6 : \ (x \vee y) \to z = (x \to z) \wedge (y \to z)$ and $I_7 : \ (x \wedge y) \to z = (x \to z) \vee (y \to z)$.

(a) Let $x = v_{i1}, y =_{j1}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6 : \ (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v_{\min(n, n-i+k)1}, \ if \ i \geq j \\ v_{\min(n, n-j+k)1}, \ if \ i < j \end{cases}$$

$$I_7 : \ (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v_{\min(n, n-j+k)1}, \ if \ i \geq j \\ v_{\min(n, n-i+k)1}, \ if \ i \leq j \end{cases}$$

(b) Let $x = v_{i1}, y =_{j1}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6: (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v\max_{(0,i+k-n)0}, & if\ i \geq j \\ v\max_{(0,j+k-n)0}, & if\ i < j \end{cases}$$

$$I_7: (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v\max_{(0,j+k-n)0}, & if\ i \geq j \\ v\max_{(0,i+k-n)1}, & if\ i \leq j \end{cases}$$

(c) Let $x = v_{i1}, y =_{j0}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6: (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v\min_{(n,n-i+k)1}, & if\ i, j \geq \frac{n}{2} \\ v\min_{(n,j+k)1}, & if\ i, j < \frac{n}{2} \end{cases}$$

$$I_7: (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v\min_{(n,j+k)1}, & if\ i, j \geq \frac{n}{2} \\ v\min_{(n,n-i+k)1}, & if\ i, j < \frac{n}{2} \end{cases}$$

(d) Let $x = v_{i1}, y =_{j0}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6: (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v\max_{(0,i+k-n)0}, & if\ i, j \geq \frac{n}{2} \\ v\min_{(n,n+j-k)1}, & if\ i, j < \frac{n}{2} \end{cases}$$

$$I_7: (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v\min_{(n,n+j-k)1}, & if\ i, j \geq \frac{n}{2} \\ v\max_{(0,i+k-n)0}, & if\ i, j < \frac{n}{2} \end{cases}$$

(e) Let $x = v_{i0}, y =_{j1}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6: (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v\min_{(n,n-j+k)1}, & if\ i, j \geq \frac{n}{2} \\ v\min_{(n,j+k)1}, & if\ i, j < \frac{n}{2} \end{cases}$$

$$I_7: (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v\min_{(n,i+k)1}, & if\ i, j \geq \frac{n}{2} \\ v\min_{(n,n-i+k)1}, & if\ i, j < \frac{n}{2} \end{cases}$$

(f) Let $x = v_{i0}, y =_{j1}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6: (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v\max_{(0,j+k-n)1}, & if\ i, j \geq \frac{n}{2} \\ v\min_{(n,n+i-k)1}, & if\ i, j < \frac{n}{2} \end{cases}$$

$$I_7:\ (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v_{\min(n,n+i-k)1}, & if\ i, j \geq \frac{n}{2} \\ v_{\max(0,j+k-n)1}, & if\ i, j < \frac{n}{2} \end{cases}$$

(g) Let $x = v_{i0}, y =_{j0}, z = v_{k1}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6:\ (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v_{\min(n,j+k)1}, & if\ i \geq j \\ v_{\min(n,i+k)1}, & if\ i < j \end{cases}$$

$$I_7:\ (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v_{\min(n,i+k)1}, & if\ i \geq j \\ v_{\min(n,j+k)1}, & if\ i \leq j \end{cases}$$

(h) Let $x = v_{i0}, y =_{j0}, z = v_{k0}, \forall i, j, k \in \{0, 1, 2, \ldots, n\}$

$$I_6:\ (x \vee y) \to z = (x \to z) \wedge (y \to z) = \begin{cases} v_{\min(n,n-k+j)1}, & if\ i \geq j \\ v_{\min(n,n-k+i)1}, & if\ i < j \end{cases}$$

$$I_7:\ (x \wedge y) \to z = (x \to z) \vee (y \to z) = \begin{cases} v_{\min(n,n-k+i)1}, & if\ i \geq j \\ v_{\min(n,n-k+j)1}, & if\ i \leq j \end{cases}$$

Thus, L is a *lattice implication algebra.*

Definition 2.3 Propositions having linguistic truth values are called Linguistic Truth-valued Propositions (LTVP) and is denoted by $\Im$. If e is a truth value of a LTVP then $e : \Im \to L$.

Definition 2.4 Let $P \in \Im$ denote an atom, the fundamental element of LTVP. Any formula of LTVP is defined recursively as follows:

1. P is a formula.
2. If G is a formula then G' is also a formula.
3. If G, H are LTVP formulae then $G \wedge H$, $G \vee H$, $G \to H$ are formulae.
4. Any symbolic string formed using 1, 2, 3 a finite number of times is called a formula in LTVP. No other string is a formula.

Definition 2.5 Truth value of a formula of LTVP is defined recursively as folows:

1. If $G \in \Im$ then $e(G') = (e(G))'$ is also a formula.
2. If $G, H \in \Im$ then $e(G \wedge H) = e(G) \wedge e(H)$; $e(G \vee H) = e(G) \vee e(H)$ and $e(G \to H) = e(G) \to e(H)$.

Next we discuss the case where some of the elements of V do not satisfy the following order $i \in \{0, 1, 2, \ldots, n\}$, $v_{i0} \leq v_{(n-i)1}$, i.e. $\exists$ at least one i for which the elements v_{i0}, $v_{(n-i)1}$ are non- comparable, then also V is a partially ordered set and may be represented by the Hasse diagram given by Fig. 2.

Theorem 2.4 If V be the linguistic truth value set, then $(V, \leq)$ is a poset. Let $\chi = (V, \wedge, \vee, O, I)$ where $O = v_{n0},\ I = v_{n1}, v_{i0}$ and $v_{(n-i)1}$ are non-comparable (Fig. 2) and $\forall\ k, l \in \{0, 1, \ldots, n\},\ r = \max\{k, l\},\ s = \min\{k, l\}$. If the operation "$\wedge$"and "$\vee$"are defined as:

1. $v_{k1} \vee v_{l1} = v_{r1}$
2. $v_{k0} \vee v_{l0} = v_{s0}$
3. $v_{k1} \vee v_{l0} = \begin{cases} v_{k1}, & k \neq (n-i) \\ v_{(n-(i-1))1}, & k = (n-i) \end{cases} \quad if\ n \leq (k+l)$
4. $v_{k1} \vee v_{l0} = \begin{cases} v_{(n-l)1}, & l \neq i \\ v_{(n-(i-1))1}, & l = i \end{cases} \quad if\ n \geq (k+l)$
5. $v_{k1} \wedge v_{l1} = v_{s1}$
6. $v_{k0} \wedge v_{l0} = v_{r0}$
7. $v_{k1} \wedge v_{l0} = \begin{cases} v_{l0}, & l \neq i \\ v_{(i+1)0}, & k = (n-i),\ l = i \end{cases} \quad if\ n \leq (k+l)$
8. $v_{k1} \wedge v_{l0} = \begin{cases} v_{(n-k)0}, & k \neq (n-i) \\ v_{(i+1)0}, & k = (n-i) \end{cases} \quad if\ n \geq (k+l)$

Proof The proof is similar to that of the Theorem 2.2.

With the unary operation (inverse operation) "$'$" and the binary operation (implication operation) "$\rightarrow$" defined on V as given in the Definition 2.2 it may be proved that $L_1 = (V, \wedge, \vee, ', O, I, \rightarrow)$ is a Quasi-lattice implication algebra.

Theorem 2.5 *If $L_1 = (V, \wedge, \vee, ', O, I, \rightarrow)$ then L_1 is a Quasi-lattice implication algebra.*

Proof $(V, \wedge, \vee, ', O, I, \rightarrow)$ forms a bounded lattice due to Theorem 2.4 "$'$" is an order reversing involution as $\forall i \in \{1, 2, \ldots, n\},\ \forall j \in \{0, 1\};\ (v'_{ij})' = v_{ij}$.

The proof is similar to the Theorem 2.3 for the properties $I_1 - I_5$. Now we check for the properties $I_6:\ (x \vee y) \rightarrow z = (x \rightarrow z) \wedge (y \rightarrow z)$ and $I_7:\ (x \wedge y) \rightarrow z = (x \rightarrow z) \vee (y \rightarrow z)$.

Let $x = v_{(n-i)1},\ y = v_{i0},\ z = v_{k1}$, then for any $i \in \{1, 2, \ldots, (n-1)\}$ and $k \in \{0, 1, 2, \ldots, n\}$ with $i + k + 1 < n$

For property $I_6:\ (x \vee y) \rightarrow z = v_{(i+k-1)1}\ and\ (x \rightarrow z) \wedge (y \rightarrow z) = v_{(i+k)1}$

For property $I_7:\ (x \wedge y) \rightarrow z = v_{(i+k+1)1}\ and\ (x \rightarrow z) \vee (y \rightarrow z) = v_{(i+k)1}$

Thus, we see that L_1 does not satisfy the properties I_6 and I_7 and therefore it is a *Quasi-lattice implication algebra.*

Definition 2.6 Propositions having Quasi linguistic truth values are called Quasi Linguistic truth-valued Propositions (QLTVP) and is denoted by $\Re$. If e is a truth valuation of a QLTVP then $e : \Re \rightarrow L_1$.

Definition 2.7 Let $Q \in \Re$ denote an atom, the fundamental element of QLTVP. Any formula of QLTVP is defined recursively as follows:

1. Q is a formula.

2. If G is a formula then G' is also a formula.
3. If G, H are QLTVP formulae then $G \wedge H,\ G \vee H,\ G \to H$ are formulae.
4. Any symbolic string formed using 1, 2, 3 a finite number of times is called a formula in QLTVP. No other string is a formula.

3 Reasoning with Linguistic Truth-Valued Propositions

Truth values of Modus Ponens and Modus Tollens rules for Linguistic truth-valued propositions (LTVP) and Quasi linguistic truth-valued propositions (QLTVP) are computed in this section. We assume that P and Q are two propositions having linguistic truth values $e(P),\ e(Q) \in V$ and $e(P \to Q) = v_{j1}$ where $j \in \{0, 1, \ldots, n\}$. With this assumption we observed that the fundamental reasoning tools are not always absolutely true for LTVP.

3.1 Computation of Truth Values of Modus Ponens and Modus Tollens rules for LTVP

Let $= \{P|P$ is a LTVP$\}$ be the set of all LTVPs. The truth evaluation e is a function given by $e : \Im \to V$. Henceforth we will denote P' by $\neg P$.

Theorem 3.1.1 *If* $P, Q \in \Im$ *and* $e(P) = v_{i1};\ e(P \to Q) = v_{j1}$ *and* $e(Q) = v_{k1}$ *then the truth values of the Modus Ponens and Modus Tollens rules are*

$$e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ i \le j,\ i \le k \\ v_{n1}, & if\ i \ge j, j \le k \\ v_{(n-i+k)1}, & if\ i \le j, i \ge k \\ v_{(n-j+k)1}, & if\ i \ge j, j \ge k \end{cases}$$

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ n \le j+k,\ i \le k \\ v_{n1}, & if\ i+j \le n, j+k \le n \\ v_{(n-i+k)1}, & if\ n \le j+k, i \ge k \\ v_{(2n-i-j)1}, & if\ i+j \ge n, j+k \le n \end{cases}$$

Proof Case 1: Modus Ponens Rule:

$$e(P \wedge (P \to Q)) = \begin{cases} v_{i1}, & if\ i \le j \\ v_{j1}, & if\ i \ge j \end{cases}$$

So, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{\min(n,n-i+k)1}, & if\ i \le j \\ v_{\min(n,n-j+k)1}, & if\ i \ge j \end{cases}$

Therefore, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ i \le j,\ i \le k \\ v_{n1}, & if\ i \ge j, j \le k \\ v_{(n-i+k)1}, & if\ i \le j, \ge i \ge k \\ v_{(n-j+k)1}, & if\ i \ge j, j \ge k \end{cases}$

Case 2: Modus Tollens Rule:

$$e(\neg Q \wedge (P \to Q)) = \begin{cases} v_{k0}, & if\ n \le k + j \\ v_{(n-j)0}, & if\ n \ge k + j \end{cases}$$

So, $e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{\min(n,n-i+k)1}, & if\ n \le k + j \\ v_{\min(n,2n-i-j)}, & if\ n \ge k + j \end{cases}$

Therefore, $e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ n \le j + k,\ i \le k \\ v_{n1}, & if\ i + j \le n, j + k \le n \\ v_{(n-i+k)1}, & if\ n \le j + k, i \ge k \\ v_{(2n-i-j)1}, & if\ i + j \ge n, j + k \le n \end{cases}$

Remark Modus Ponens rule is absolutely true provided for $e(P) = v_{i1};\ e(P \to Q) = v_{j1}\ and\ e(Q) = v_{k1}$ either $i \le j,\ i \le k\ or\ i \ge j,\ j \le k.$

In the same case Modus Tollens rule will be absolutely true if either $n \le j+k,\ i \le k\ or\ i+j \le n,\ j+k \le n$. In all other cases the rules though true will not be exactly so.

Theorem 3.1.2 *If* $P, Q \in \Im$ *and* $e(P) = v_{i1};\ e(P \to Q) = v_{j1}$ *and* $e(Q) = v_{k0}$ *then the truth values of the Modus Ponens and Modus Tollens rules are*

$$e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{00}, & if\ i \le j,\ i + k \le n \\ v_{00}, & if\ i \ge j, j + k \le n \\ v_{(i+k-n)0}, & if\ i \le j, i + k \ge n \\ v_{(j+k-n)0}, & if\ i \ge j, j + k \ge n \end{cases}$$

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{00}, & if\ i + k \le n,\ k \le j \\ v_{00}, & if\ i + j \le n, j \le k \\ v_{(i+k-n)0}, & if\ i + k \ge n,\ k \le j \\ v_{(i+j-n)0}, & if\ i + j \ge n, j \le k \end{cases}$$

Proof Case 1: Modus Ponens Rule:

$$e(P \wedge (P \to Q)) = \begin{cases} v_{i1}, & if\ i \le j \\ v_{j1}, & if\ i \ge j \end{cases}$$

So, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{\max(0,i+k-n)0}, & if\ i \le j \\ v_{\max(0,j+k-n)0}, & if\ i \ge j \end{cases}$

Therefore, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{00}, & if\ i \le j,\ i+k \le n \\ v_{00}, & if\ i \ge j, j+k \le n \\ v_{(i+k-n)0}, & if\ i \le j, i+k \ge n \\ v_{(j+k-n)0}, & if\ i \ge j, j+k \ge n \end{cases}$

Case 2: Modus Tollens Rule:

$$e(\neg Q \wedge (P \to Q)) = \begin{cases} v_{k1}, & if\ k \le j \\ v_{j1}, & if\ k \ge j \end{cases}$$

So, $e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{\max(0,i+k-n)0}, & if\ k \le j \\ v_{\max(0,i+j-n)0}, & if\ k \ge j \end{cases}$

Therefore, $e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{00}, & if\ i+k \le n,\ k \le j \\ v_{00}, & if\ i+j \le n, j \le k \\ v_{(i+k-n)0}, & if\ i+k \ge n,\ k \le j \\ v_{(i+j-n)0}, & if\ i+j \ge n, j \le k \end{cases}$

Remark If $e(P) = v_{i1}$ whereas $e(Q) = v_{k0}$ then neither the Modus Ponens rule nor the Modus Tollens rule will be even slightly true.

Theorem 3.1.3 *If and $e(P) = v_{i0}$; $e(P \to Q) = v_{j1}$ and $e(Q) = v_{k1}$ then the truth values of the Modus Ponens and Modus Tollens rules are*

$$e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ n \le i+j,\ n \le i+k \\ v_{n1}, & if\ n \ge i+j, j \le k \\ v_{(i+k)1}, & if\ n \le i+j,\ n \ge i+k \\ v_{(n-j+k)1}, & if\ n \ge i+j, j \ge k \end{cases}$$

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ n \le i+k,\ n \le j+k \\ v_{n1}, & if\ n \ge j+k, i \ge j \\ v_{(i+k)1}, & if\ n \ge i+k,\ n \le j+k \\ v_{(n-j+i)1}, & if\ n \ge j+k, i \le j \end{cases}$$

Proof Case 1: Modus Ponens Rule:

$$e(P \wedge (P \to Q)) = \begin{cases} v_{i0}, & if\ n \le i+j \\ v_{(n-j)0}, & if\ n \ge i+j \end{cases}$$

So, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{\min(n,i+k)1}, & if\ n \le i+j \\ v_{\min(n,n-j+k)1}, & if\ n \ge i+j \end{cases}$

Therefore, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ n \le i+j,\ n \le i+k \\ v_{n1}, & if\ n \ge i+j, j \le k \\ v_{(i+k)1}, & if\ n \le i+j,\ n \ge i+k \\ v_{(n-j+k)1}, & if\ n \ge i+j, j \ge k \end{cases}$

Case 2: Modus Tollens Rule:

$$e(\neg Q \wedge (P \to Q)) = \begin{cases} v_{k0}, & if\ n \le j+k \\ v_{(n-j)0}, & if\ n \ge j+k \end{cases}$$

So, $e(\neg Q \wedge (P \to Q)\neg P) = \begin{cases} v_{\min(n,i+k)1}, & if\ n \le j+k \\ v_{\min(n,n-j+i)1}, & if\ n \ge j+k \end{cases}$

Therefore, $e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ n \le i+k,\ n \le j+k \\ v_{n1}, & if\ n \ge j+k, i \ge j \\ v_{(i+k)1}, & if\ n \ge i+k,\ n \le j+k \\ v_{(n-j+i)1}, & if\ n \ge j+k, i \le j \end{cases}$

Theorem 3.1.4 *If $P, Q \in \Im$ and $e(P) = v_{i0}$; $e(P \to Q) = v_{j1}$ and $e(Q) = v_{k0}$ then the truth values of the Modus Ponens and Modus Tollens rules are*

$$e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ n \le i+j,\ i \ge k \\ v_{n1}, & if\ n \ge i+j, n \ge j+k \\ v_{(n+i-k)1}, & if\ n \le i+j,\ i \le k \\ v_{(2n-j-k)1}, & if\ n \ge i+j, n \le j+k \end{cases}$$

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ k \le j,\ k \le i \\ v_{n1}, & if\ k \ge j, i \ge j \\ v_{(n+i-k)1}, & if\ k \ge i,\ k \le j \\ v_{(n-j+i)1}, & if\ k \ge j, i \le j \end{cases}$$

Proof Case 1: Modus Ponens Rule:

$$e(P \wedge (P \to Q)) = \begin{cases} v_{i0}, & if\ n \le i+j \\ v_{(n-j)0}, & if\ n \ge i+j \end{cases}$$

So, $e(P \wedge (P \to Q)) = \begin{cases} v_{\min(n,n+i-k)1}, & if\ n \le i+j \\ v_{\min(n,2n-j-k)1}, & if\ n \ge i+j \end{cases}$

Therefore, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ n \le i+j,\ i \ge k \\ v_{n1}, & if\ n \ge i+j, n \ge j+k \\ v_{(n+i-k)1}, & if\ n \le i+j,\ i \le k \\ v_{(2n-j-k)1}, & if\ n \ge i+j, n \le j+k \end{cases}$

Case 2: Modus Tollens Rule:

$$e(\neg Q \wedge (P \to Q)) = \begin{cases} v_{k1}, & if\ k \le j \\ v_{j1}, & if\ k \ge j \end{cases}$$

So, $e(\neg Q \wedge (P \to Q)\neg P) = \begin{cases} v_{\min(n,n+i-k)1}, & if\ k \le j \\ v_{\min(n,n-j+i)1}, & if\ k \ge j \end{cases}$

$$\text{Therefore, } e(\neg Q \wedge (P \rightarrow Q) \rightarrow \neg P) = \begin{cases} v_{n1}, & if\ k \leq j,\ k \leq i \\ v_{n1}, & if\ k \geq j, i \geq j \\ v_{(n+i-k)1}, & if\ k \geq i,\ k \leq j \\ v_{(n-j+i)1}, & if\ k \geq j, i \leq j \end{cases}$$

3.2 Computation of Truth Values of Modus Ponens and Modus Tollens rules for QLTVP

Truth values of Modus Ponens and Modus Tollens rules for *Quasi-Linguistic truth-valued Propositions (QLTVP)* are computed in this section. It is observed that these fundamental reasoning tools are not always absolutely true.

In this section V, the set of linguistic truth values have a lattice structure represented by Fig. 3 (i.e. v_{l0} and $v_{(n-l)1}$ are non-comparable). Let $\Re = \{Q|Q$ is a QLTVP$\}$be the set of all QLTVPs. The truth evaluation e is a function given by $e : \Re \rightarrow V$.

Theorem 3.2.1 *If* $P, Q \in \Re$ *and* $e(P) = v_{i1}$; $e(P \rightarrow Q) = v_{j1}$ *and* $e(Q) = v_{k1}$ *then the truth values of the Modus Ponens and Modus Tollens rules are*

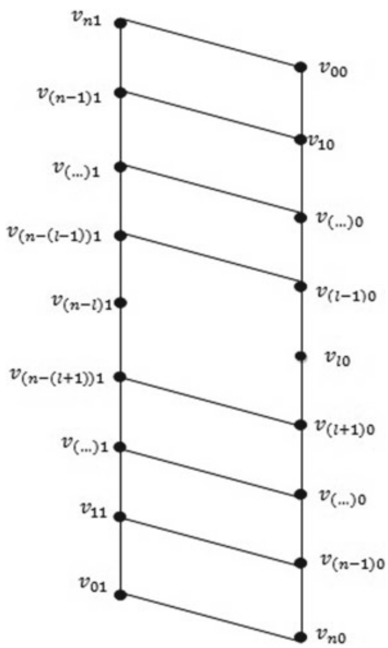


Fig. 3 Hasse diagram of linguistic truth value set V satisfying QLIA

$$e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ i \le j,\ i \le k \\ v_{n1}, & if\ i \ge j, j \le k \\ v_{(n-i+k)1}, & if\ i \le j, i \ge k \\ v_{(n-j+k)1}, & if\ i \ge j, j \ge k \end{cases}$$

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ n \le j+k, i \le k, k \ne l \\ v_{n1}, & if\ i \le l+1, j = n-l \\ v_{n1}, & if\ i+j \le n, j+k \le n, j \ne n-l \\ v_{(n-i+k)1}, & if\ n \le j+k, i \ge k, k \ne l \\ v_{(n-i+l+1)1}, & if\ i \ge l+1, j = n-l \\ v_{(2n-i-j)1}, & if\ i+j \ge n, j+k \le n, j \ne n-l \end{cases}$$

Proof Case 1: Modus Ponens Rule:

$$e(P \wedge (P \to Q)) = \begin{cases} v_{i1}, & if\ i \le j \\ v_{j1}, & if\ i \ge j \end{cases}$$

So, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{\min(n,n-i+k)1}, & if\ i \le j \\ v_{\min(n,n-j+k)1}, & if\ i \ge j \end{cases}$

Therefore, $e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ i \le j,\ i \le k \\ v_{n1}, & if\ i \ge j, j \le k \\ v_{(n-i+k)1}, & if\ i \le j, i \ge k \\ v_{(n-j+k)1}, & if\ i \ge j, j \ge k \end{cases}$

Case 2: Modus Tollens Rule:

$$e(\neg Q \wedge (P \to Q)) = \begin{cases} v_{k0}, & if\ n \le k+j, k \ne l \\ v_{(l+1)0}, & if\ n \le j+k, k = l, j = n-l \\ v_{(n-j)0}, & if\ n \ge j+k, j \ne n-l \\ v_{(l+1)0}, & if\ n \ge j+k, j = n-l \end{cases}$$

So,

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{\min(n,n-i+k)1}, & if\ n \le k+j, k \ne l \\ v_{\min(n,n-i+l+1)1}, & if\ n \le j+k, k = l, j = n-l \\ v_{\min(n,2n-i-j)1}, & if\ n \ge j+k, j \ne n-l \\ v_{\min(n,n-i+l+1)1}, & if\ n \ge j+k, j = n-l \end{cases}$$

Therefore, $e(\neg Q \wedge (P \rightarrow Q) \rightarrow \neg P) = \begin{cases} v_{n1}, & if\ n \le j+k, i \le k, k \ne l \\ v_{n1}, & if\ i \le l+1,\ j = n-l \\ v_{n1}, & if\ i+j \le n,\ j+k \le n,\ j \ne n-l \\ v_{(n-i+k)1}, & if\ n \le j+k, i \ge k, k \ne l \\ v_{(n-i+l+1)1}, & if\ i \ge l+1,\ j = n-l \\ v_{(2n-i-j)1}, & if\ i+j \ge n,\ j+k \le n,\ j \ne n-l \end{cases}$

Theorem 3.2.2 *If $P, Q \in \Re$ and $e(P) = v_{i1}$; $e(P \rightarrow Q) = v_{j1}$ and $e(Q) = v_{k1}$ then the truth values of the Modus Ponens and Modus Tollens rules are*

$$e(P \wedge (P \rightarrow Q) \rightarrow Q) = \begin{cases} v_{n1}, & if\ n \le i+j, n \le i+k, i \ne l \\ v_{n1}, & if\ n \le k+l+1,\ j = n-l \\ v_{n1}, & if\ i+j \le n,\ j \le k,\ j \ne n-l \\ v_{(i+k)1}, & if\ n \le i+j, n \ge i+k, i \ne l \\ v_{(k+l+1)1}, & if\ n \ge k+l+1,\ j = n-l \\ v_{(n-j+k)1}, & if\ i+j \le n,\ j \ge k,\ j \ne n-l \end{cases}$$

$$e(\neg Q \wedge (P \rightarrow Q) \rightarrow \neg P) = \begin{cases} v_{n1}, & if\ n \le k+j, n \le i+k, k \ne l \\ v_{n1}, & if\ n \le i+l+1,\ j = n-l \\ v_{n1}, & if\ k+j \le n,\ j \le i,\ j \ne n-l \\ v_{(i+k)1}, & if\ n \le k+j, n \ge i+k, k \ne l \\ v_{(i+l+1)1}, & if\ n \ge i+l+1,\ j = n-l \\ v_{(n-j+i)1}, & if\ k+j \le n,\ j \ge i,\ j \ne n-l \end{cases}$$

Proof Case 1: Modus Ponens Rule:

$$e(P \wedge (P \rightarrow Q)) = \begin{cases} v_{i0}, & if\ n \le i+j, i \ne l \\ v_{(l+1)0}, & if\ n \le j+i, i = l,\ j = n-l \\ v_{(n-j)0}, & if\ n \ge j+i,\ j \ne n-l \\ v_{(l+1)0}, & if\ n \ge j+i,\ j = n-l \end{cases}$$

$$\text{So}, e(P \wedge (P \rightarrow Q) \rightarrow Q) = \begin{cases} v_{\min(n,i+k)1}, & if\ n \le i+j, i \ne l \\ v_{\min(n,k+l+1)1}, & if\ n \le j+i, i = l,\ j = n-l \\ v_{\min(n,n-j+k)1}, & if\ n \ge j+i,\ j \ne n-l \\ v_{\min(n,k+l+1)1}, & if\ n \ge j+i,\ j = n-l \end{cases}$$

Therefore, $e(P \wedge (P \to Q) \to Q) =$
$$\begin{cases} v_{n1}, & if\ n \le i+j, n \le i+k, i \ne l \\ v_{n1}, & if\ n \le k+l+1, j = n-l \\ v_{n1}, & if\ i+j \le n, j \le k, j \ne n-l \\ v_{(i+k)1}, & if\ n \le i+j, n \ge i+k, i \ne l \\ v_{(k+l+1)1}, & if\ n \ge k+l+1, j = n-l \\ v_{(n-j+k)1}, & if\ i+j \le n, j \ge k, j \ne n-l \end{cases}$$

Case 2: Modus Tollens Rule:

$$e(\neg Q \wedge (P \to Q)) = \begin{cases} v_{i0}, & if\ n \le k+j, k \ne l \\ v_{(l+1)0}, & if\ n \le j+k, k = l, j = n-l \\ v_{(n-j)0}, & if\ n \ge j+k, j \ne n-l \\ v_{(l+1)0}, & if\ n \ge j+k, j = n-l \end{cases}$$

So,

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{\min(n,i+k)1}, & if\ n \le k+j, k \ne l \\ v_{\min(n,i+l+1)1}, & if\ n \le j+k, k = l, j = n-l \\ v_{\min(n,n-j+i)1}, & if\ n \ge j+k, j \ne n-l \\ v_{\min(n,i+l+1)1}, & if\ n \ge j+k, j = n-l \end{cases}$$

Therefore, $e(\neg Q \wedge (P \to Q) \to \neg P) =$
$$\begin{cases} v_{n1}, & if\ n \le k+j, n \le i+k, k \ne l \\ v_{n1}, & if\ n \le i+l+1, j = n-l \\ v_{n1}, & if\ k+j \le n, j \le i, j \ne n-l \\ v_{(i+k)1}, & if\ n \le k+j, n \ge i+k, k \ne l \\ v_{(i+l+1)1}, & if\ n \ge i+l+1, j = n-l \\ v_{(n-j+i)1}, & if\ k+j \le n, j \ge i, j \ne n-l \end{cases}$$

Theorem 3.2.3 *If $P, Q \in \Re$ and $e(P) = v_{i1}$; $e(P \to Q) = v_{j1}$ and $e(Q) = v_{k1}$ then the truth values of the Modus Ponens and Modus Tollens rules are*

$$e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{n1}, & if\ n \le i+j, i \ge k, i \ne l \\ v_{n1}, & if\ k \le l+1, j = n-l \\ v_{n1}, & if\ i+j \le n, j+k \le n, j \ne n-l \\ v_{(n+i-k)1}, & if\ n \le i+j, i \le k, i \ne l \\ v_{(n-k+l+1)1}, & if\ k \ge l+1, j = n-l \\ v_{(2n-j-k)1}, & if\ i+j \le n, j+k \ge n, j \ne n-l \end{cases}$$

$$e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ k \le j,\ k \le i \\ v_{n1}, & if\ k \ge j, j \le i \\ v_{(n-k+i)1}, & if\ k \le j, k \ge i \\ v_{(n-j+i)1}, & if\ k \ge j, j \ge i \end{cases}$$

Proof Case 1: Modus Ponens Rule:

$$e(P \wedge (P \to Q)) = \begin{cases} v_{i0}, & if\ n \le i+j, i \ne l \\ v_{(l+1)0}, & if\ n \le j+i, i = l, j = n-l \\ v_{(n-j)0}, & if\ n \ge j+i, j \ne n-l \\ v_{(l+1)0}, & if\ n \ge j+i, j = n-l \end{cases}$$

So,

$$e(P \wedge (P \to Q) \to Q) = \begin{cases} v_{\min(n,n+i-k)1}, & if\ n \le i+j, i \ne l \\ v_{\min(n,n-k+l+1)1}, & if\ n \le j+i, i = l, j = n-l \\ v_{\min(n,2n-j-k)1}, & if\ n \ge j+i, j \ne n-l \\ v_{\min(n,n-k+l+1)1}, & if\ n \ge j+i, j = n-l \end{cases}$$

Therefore, $e(P \wedge (P \to Q) \to Q) =$
$$\begin{cases} v_{n1}, & if\ n \le i+j, i \ge k, i \ne l \\ v_{n1}, & if\ k \le l+1, j = n-l \\ v_{n1}, & if\ i+j \le n, j+k \le n, j \ne n-l \\ v_{(n+i-k)1}, & if\ n \le i+j, i \le k, i \ne l \\ v_{(n-k+l+1)1}, & if\ k \ge l+1, j = n-l \\ v_{(2n-j-k)1}, & if\ i+j \le n, j+k \ge n, j \ne n-l \end{cases}$$

Case 2: Modus Tollens Rule:

$$e(\neg Q \wedge (P \to Q)) = \begin{cases} v_{k1}, & if\ k \le j \\ v_{j1}, & if\ k \ge j \end{cases}$$

So, $e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{\min(n,n-k+i)1}, & if\ k \le j \\ v_{\min(n,n-j+i)1}, & if\ k \ge j \end{cases}$

Therefore, $e(\neg Q \wedge (P \to Q) \to \neg P) = \begin{cases} v_{n1}, & if\ k \le j,\ k \le i \\ v_{n1}, & if\ k \ge j, j \le i \\ v_{(n-k+i)1}, & if\ k \le j, k \ge i \\ v_{(n-j+i)1}, & if\ k \ge j, j \ge i \end{cases}$

4 Output and Results

In real life we very often come across situations where we have to reason with linguistic truth valued or quasi-linguistic truth valued propositions. Hence there arises a problem of validation of the argument given. Let A be the logical expression representing the argument. We have computed the truth value of A $(e(A))$ in different cases and found that the argument may be *Absolutely true, Quite true,……Slightly true, Slightly false,……Quite false, Absolutely false.* Given the truth value of the premise and the consequent by the respective formula, the truth of the validity of the inference rule may be computed.

We usually try to take a decision whether the premise implies the consequent. If the consequent has a truth value somewhat true and the computed value of the argument is quite true then we may say that *Given the premise it is quite true that the premise implies that the consequent is somewhat true.* Since quite true is a high truth value so we may take the decision that “The consequent is somewhat true”. However if the computed value is slightly true/slightly false we may reject the decision and if the computed value is absolutely false we say “The consequent is not at all somewhat true”. We use the above method in the following examples. The lattice structure is represented by Fig. 4.

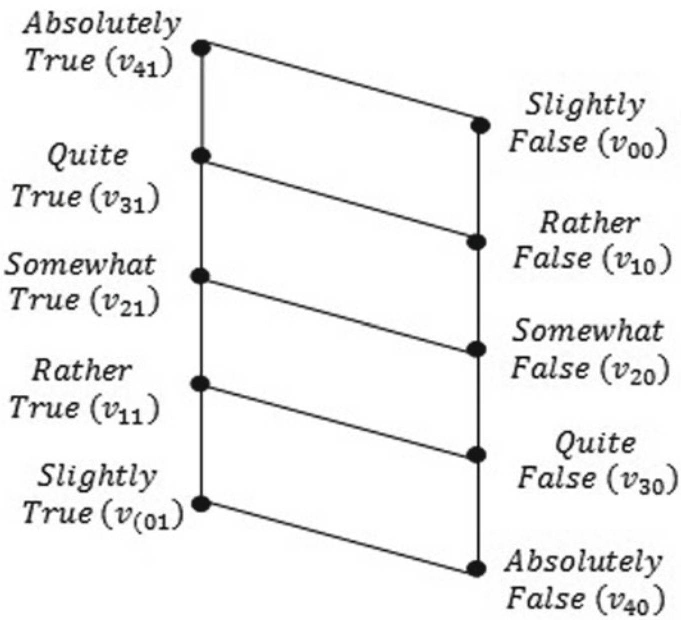


Fig. 4 Hasse diagram of linguistic truth value set V satisfying LIA having some specific truth-values

4.1 Examples

Example 4.1.1 (*Case-5* of the data sheet given in Table 2 of [9]).

A 66 yr. old patient (case-5) had fever, cough, fatigue and headache as symptoms and pulmonary disease, hypertension, diabetes, chronic renal insufficiency as pre-existing health problems then it is absolutely true that he will die due to COVID-19.

Solution:

Let, A: The patient is old.

S: The patient has four symptoms of COVID-19.

H: The patient has four pre-existing health problems.

Q: The patient died due to COVID-19.

Let $P = (A \wedge S) \wedge H$.

Since the patient had quite high symptoms and co morbidity so $e(P \to Q) = v_{31}/v_{41}$.

Also $e(A) = v_{41};\ e(S) = v_{31};\ e(H) = v_{31}$.

Therefore, $e(P) = v_{31}$ and $e(Q) = v_{41}$.

Thus, Modus Ponens rule is

$$e(P \wedge (P \to Q) \to Q) = v_{n1} = v_{41}.(\because\ i \leq j\ and\ i < k)$$

and Modus Tollens rule is

$$e(\neg Q \wedge (P \to Q) \to \neg P) = v_{n1} = v_{41}.(\because\ i < k\ and\ n < j + k)$$

So it is absolutely true that the given premise imply that the patient died due to Covid 19. The table confirms the result.

Example 4.1.2 (*Case-12* of the data sheet given in Table 2 of [9]).

A 65 yr. old patient (case-12) had anhelation and fatigue and no pre-existing health problems then it is slightly true that he will die due to COVID-19.

Solution:

Let, A: The patient is old.

S: The patient has two symptoms of COVID-19.

H: The patient has no pre-existing health problems.

Q: The patient died due to COVID-19.

Let $P = (A \wedge S) \wedge H$.

Since the patient had rather high symptoms, so $e(P \to Q) = v_{41}$;

Also $e(A) = v_{41};\ e(S) = v_{11};\ e(H) = v_{00}$.

Therefore, $e(P) = v_{30}$ and $e(Q) = v_{01}$.

Thus, Modus Ponens rule is

$$e(P \wedge (P \to Q) \to Q) = v_{(i+k)1} = v_{31}.(\because\ j + k > n\ and\ i + k < n)$$

and Modus Tollens rule is

$$e(\neg Q \wedge (P \rightarrow Q) \rightarrow \neg P) = v_{(n-j+i)1} = v_{31}.(\because\ n = j + k \text{ and } i < j)$$

So it is quite true that the given premise imply that the death of the patient due to Covid 19 is slightly true. The table confirms the result.

Example 4.1.3 (*Case-3* of the data sheet given in Table 3 [17]).

If a patient is between the age 70–79, had slight temperature, showed symptoms of pneumonia may have been in contact with a COVID-19 positive patient then it is slightly true that he will get infected by COVID-19.

Solution:

Let, A: The patient is old.

S: The patient has two symptoms of COVID-19.

C: The patient may have come in contact with any COVID-19 positive patient.

Q: The patient got infected by COVID-19.

Let $P = (A \wedge S) \wedge C$.

Since the patient had rather high symptoms, so $e(P \rightarrow Q) = v_{41}$.

Also $e(A) = v_{41};\ e(S) = v_{21};\ e(C) = v_{21}$.

Therefore, $e(P) = v_{21}$ and $e(Q) = v_{01}$.

Thus, Modus Ponens rule is

$$e(P \wedge (P \rightarrow Q) \rightarrow Q) = v_{(n-i+k)1} = v_{21}.(\because\ i < j \text{ and } i > k)$$

and Modus Tollens rule is

$$e(\neg Q \wedge (P \rightarrow Q) \rightarrow \neg P) = v_{(n-i+k)1} = v_{21}.(\because\ n = j + k \text{ and } i > k)$$

So it is somewhat true that the given premise imply that the possibility of the patient being infected by COVID 19 is slightly true. The table confirms the result.

Example 4.1.4 (*Case-4* of the data sheet given in Table 3 [17]).

If a patient is in the age group 40–49 and has fever but nothing is known about co morbidity then it is somewhat true that he will get infected from COVID-19.

Solution:

Let, A: The patient is old.

S: The patient has one symptom of COVID-19.

Q: The patient got infected by COVID-19.

Let $P = (A \wedge S)$.

Here, $e(P \rightarrow Q) = v_{41}$.

Also $e(A) = v_{21};\ e(S) = v_{11}$.

Therefore, $e(P) = v_{11}$ and $e(Q) = v_{21}$.

Thus, Modus Ponens rule is

$$e(P \wedge (P \to Q) \to Q) = v_{n1} = v_{41}.(\because\ i < j\ and\ i \leq k)$$

and Modus Tollens rule is

$$e(\neg Q \wedge (P \to Q) \to \neg P) = v_{n1} = v_{41}.(\because\ n \leq j + k\ and\ i \leq k)$$

So it is absolutely true that the given premise imply that it is somewhat true that the patient got infected by COVID 19. The table confirmed that the patient was treated at home.

5 Conclusion

In this chapter we have restricted ourselves to computation of truth values of Modus Ponens and Modus Tollens rule with LTVP/QLTVP statements. For similar statements computation of truth values of Syllogism rule may be a worthwhile exercise.

References

1. Cock, M.E., Kerre, E.E.: A context-based approach to linguistic hedges. Int. J. Appl. Math. Comput. Sci. **12**(3), 371382 (2002)
2. Di Lascio, L., Gisolfi, A., Loia, V.: A new model for linguistic modifiers. Int. J. Approx. Reason. **15**, 25–47 (1996)
3. Huynh, V.N., Ho, T.B, Nakamori, Y.: A parametric representation of lin-guistic hedges in Zadeh's fuzzy logic. Int. J. Approx. Reason. **30**, 203–223 (2002)
4. Lukasiewicz, J.: O logice trjwartociowej (in Polish). Ruch filozoficzny 5: 170–171, 1920. English translation: On three-valued logic. In: Borkowski, L. (ed.) Selected Works by Jan Lukasiewicz, pp. 87–88. NorthHolland, Amsterdam (1970)
5. Nguyen, C.H., Huynh, V.N.: An algebraic approach to linguistic hedges in Zadehs fuzzy logic. Fuzzy Sets Syst. **129**, 229–254 (2002)
6. Nguyen, T.M., Khanh, T.D.: Linguistic-valued logics based on Hedge algebras and applications to approximate reasoning. Appl. Math. Inf. Sci. **11**(5), 1317–1334 (2017)
7. Nguyen, C.H., Wechler, W.: Hedge algebras: an algebraic approach to structure of sets of linguistic truth values. Fuzzy Sets Syst. **35**, 281–293 (1990)
8. Nguyen, C.H., Wechler, W.: Extended hedge algebras and their application to fuzzy logic. Fuzzy Sets Syst. **52**, 259–281 (1992)
9. Wang, W., Tang, J., Wei, F.: Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. J. Med. Virol. **92**, 441–447 (2020)
10. Xu, Y., Ruan, D., Qin, K., Liu, J.: Lattice-Valued Logic. Springer (2004)
11. Xu, Y., Qin, K.Y.: Lattice-value proposition logic (I). J. Southwest Jiaotong Univ. **12**(2), 123–128 (1993)
12. Yiquan, Z., Wenbiao, T.: A note on lattice-implication algebras. Bull. Korean Math. Soc. **38**(1), 191–195 (2001)
13. Zadeh, L.A.: The concept of linguistic variable and its application toapproximate reasoning (I). Inf. Sci. **8**, 199–249 (1975)

14. Zadeh, L.A.: The concept of linguistic variable and its application toapproximate reasoning (II). Inf. Sci. **8**, 310–357 (1975)
15. Zadeh, L.A.: The concept of linguistic variable and its application toapproximate reasoning (III). Inf. Sci. **9**, 43–80 (1975)
16. Zou, L., Ma, J., Xu, Y.: A Framework of Linguistic Truth-Valued Propositional Logic Based on Lattice Implication Algebra. IEEE (2006)
17. Worldometer for COVID-19 (2020). https://www.worldometers.info

Sumita Basu have retired as Professor and H.O.D. from Bethune College. She was associated with undergraduate teaching for 36 years and postgraduate teaching for last 14 years. At present she is working as Guest faculty in Heritage College. Her research interests include Automata theory, Mathematical logic, Rough Sets and Decision Analysis. Currently she is working on Cellular Automata, Approximate Reasoning, Variable Precision Rough Sets and Decision Analysis. She has two Ph.D. students and thirteen publications.

Bithi Chattaraj received her B.Sc. (Hons.) in Mathematics and M.Sc. in Applied Mathematics from University of Calcutta in 2009 and 2011, respectively. She qualified the GATE and NET in 2013 and got registered for Ph.D. in the department of Applied Mathematics, University of Calcutta, in 2017. She joined the department of Mathematics, Bethune College, in January, 2012 as a teacher. She has one journal publication and presented four papers in various national and international conferences. Her research interest includes Approximate Reasoning.

Analysis of Customers' Reviews Using Soft Computing Classification Algorithms: A Case Study of Amazon

Koushal Kumar and Bhagwati Prasad Pande

Abstract Recently, text and sentiment analysis has received tremendous attention, especially due to the availability of gigantic data in the form of unstructured text available on social media, E-commerce websites, E-mails, blogs, and other similar sources. It involves analyzing large volumes of unstructured text, extracting relevant information, and determining people's opinions and expressions like positive, negative, neutral etc. Nowadays, the majority of business firms are using text and sentiment analysis techniques to understand the feedbacks of their customers and to gain information about the degree of customers' inclination towards their products and services. Therefore, sentiment analysis provides valuable insights and helps the firms to formulate effective business strategies. However, the massive data derived from social media and other sources are unstructured, highly dimensional, and involve uncertainty and imprecision. Thanks to soft computing techniques, we are equipped to handle uncertainty imprecision, partial truth, and approximation. The present chapter is based on text and sentiment analysis of customers' reviews collected from the *Amazon* customer review portal. We propose a three-tier model that takes raw data from this portal as input and generates a comparative report over certain parameters. We fetch data variables from this portal, apply data preprocessing and cleaning techniques to repair and/or remove dirty data in the first phase. In the second phase, we filter out those input variables which exhibit the strongest relationship with output variables using statistical feature selection techniques. In the final phase, we analyze processed dataset using machine learning algorithms to classify positive, negative and neutral reviews. For classification, we apply Random Forest, Naïve Bayes, and Support Vector Machine algorithms in particular. These algorithms are applied to the processed dataset to study a few parameters like accuracy, precision, F-measure, true positive, false negative, etc. Finally, our study compares the outputs of these three classifiers over the above-mentioned parameters.

K. Kumar
Sikh National College, Qadian, Guru Nanak Dev University, Amritsar, Punjab, India

B. P. Pande (✉)
Department of Computer Applications, LSM Government PG College, Pithoragarh, Uttarakhand, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_15

Keywords Text and sentiment analysis · Social media · Customers' reviews · Amazon · Statistical feature selection · Machine learning · Classifiers

1 Introduction

Computing in terms of computer technology refers to the process of executing tasks with the help of a computer device. A few characteristics of the computing process are: solutions must be precise and valid, there should be unambiguous and correct control sequences, and finally, formulation of the mathematical solutions of the problems should be easy. Computing methods are categorized into two folds: hard computing and soft computing. Hard computing is the traditional practice that draws on the postulates of precision, certainty, rigour, and inflexibility. It requires a well defined analytical model and often takes a significant amount of computation time. On the other hand, soft computing is different from conventional computing, it includes the concept of approximate models and provides solutions to tricky real-world problems. Unlike hard computing, soft computing deals with imprecision, uncertainty, approximations, and partial truths. Soft computing incorporates modern theories and practices such as expert systems, fuzzy logic, genetic algorithms, artificial neural networks, and machine learning. Some of the key differences between hard and soft computing are: hard computing draws on binary (two-valued) logic and deterministic in nature, while soft computing works upon formal (multi-valued) logic and stochastic reasoning; hard computing requires exact data for its mechanism, while soft computing can tackle ambiguous and noisy data; hard computing executes sequential computations, on the other hand, soft computing is capable of performing parallel computations; hard computing requires explicit programs to be written, while soft computing can emerge its own programs.

Sentiment analysis is one of the soft computing techniques that perceives positive, negative, or neutral opinions, known as *polarity* within a piece of text. This text can be a clause, sentence, paragraph, or a whole document. Let us take customer feedback as an illustration, sentiment analysis weighs the inclination of customers towards a product or service, which they express in textual form as comments or feedbacks. For example, consider the following feedbacks by two different customers (Table 1).

The goal of sentiment analysis is to take a piece of text as input, analyze it, and returning a metric or score that estimates how positive or negative the text is. The process can be understood as context-based mining of text to identify and

Table 1 Comments and corresponding sentiments

Customer's comment (text)	Sentiment
"The watch is great! Delivery was fast, and the customer care representative was very friendly"	Positive
"I will not recommend this watch to anyone. The material is cheap, and it is truly a wastage of money"	Negative

extract subjective information from source data. It helps businesses to judge the social opinions of their products, brands, and services by monitoring the online activities of their customers. However, the analysis of web and social media platforms is limited to trivial sentiment analysis and count-based metrics: akin to engraving the surface and overlooking other important insights that ought to be discovered.

Sometimes, sentiment analysis is coupled with *text analytics* and people often consider them as the same or related processes. Though both procedures extract meaningful ideas from customer data, both are the essential constituents of the *customer experience management module*, but, they are not the same thing. As we know, the former classifies a piece of text or expression as positive, negative, or neutral and determines the degree of this classification, the latter is concerned with the analysis of the unstructured text, extracting apt information, and converting it into productive business intelligence. Text analytics deals with the *semantics* of the text: involving the grammar and the relationships among the words. In general terms, text analytics draws out the *meaning*, while sentiment analysis develops an insight into the *emotions* behind the words. Sentiment analysis has an upper hand over text analytics that the former can be applied to non-text feedbacks such as *emoticons* or *emojis*. A '*grinning face with big eyes*' emoji is coupled with a higher sentiment score than the emojis of '*frowning face*' and '*zipper-mouth face*'.

Recently, text and sentiment analysis has received tremendous attention, especially due to the availability of gigantic data in the form of unstructured text available on social media, E-commerce websites, e-mails, blogs, and other similar web resources. This requires analyzing large volumes of unstructured text, extracting relevant information, and determining people's opinions and expressions. Nowadays, the majority of business firms are using text and sentiment analysis techniques to understand the feedbacks of their customers and to gain information about the degree of customers' inclination towards their products and services. Therefore, sentiment analysis provides valuable insights and helps the firms to formulate effective business strategies. However, the massive data derived from social media and other sources are unstructured, highly dimensional, and involve uncertainty and imprecision. This kind of massive text usually contains white spaces, punctuation marks, special characters, @ links, hashtag links, stop words, and numeric digits etc. This unstructured data must be cleaned before being fed to the classification models. These types of unnecessary expressions or characters can be removed using data pre-processing libraries available in *Python*. Thanks to soft computing techniques, we are equipped to handle uncertainty imprecision, partial truth, and approximation.

The rest of the chapter is organized as follows: Sect. 2 covers the literature review, followed by Sect. 3, data collection and methodology, Sect. 4 presents experimental results and discussions, followed by the final concluding section.

2 Literature Survey

Pang and Lee [20] presented an exhaustive survey on opinion mining and sentiment analysis. They explored research works that promise to directly enable opinion-oriented information-seeking systems. They focused to give more attention to contemporary challenges raised by modern sentiment-aware applications rather than already available traditional fact-based analysis models. Prabowo and Thelwall [21] proposed a hybrid approach to sentiment analysis based on rule-based classification and machine learning. They proposed a complementary and semi-automatic approach where every classifier supports other classifiers. They tested their hybrid model over movie reviews, product reviews, and *MySpace* comments and reported that a hybrid model is capable of improving classification effectiveness in terms of micro-and macro-averaged F1 measure. The authors suggested that in real-world applications, it would be better to have two rule sets: the original and induced rule sets. Barbosa and Feng [5] investigated the writing pattern of *Twitter* messages and meta-information of the words that constitutes them. Based on this data, they proposed the automatic detection of sentiments on tweets. They utilized biased and noisy labels of tweets provided by a third party and used this source as training data. They combined these labels by utilizing various strategies and compared their model with already existing techniques. The authors claimed that the solution proposed by them can handle more abstract representation of tweets and proved to be more robust and effective. Agarwal et al. [1] studied *Twitter* data for sentiment analysis. They proposed two models: one binary model to classify tweets as positive and negative and one 3-way model to classify them as being positive, negative, and neutral sentiment. They performed experiments with the unigram model, feature-based model, and kernel-based model. The authors used the unigram model as a baseline and reported an overall gain of 4% for these classification tasks. They claimed that the feature-based and tree kernel-based models outperformed the unigram baseline. In their concluding remarks, the authors stated that the sentiment analysis for *Twitter* data is the same as sentiment analysis for other genres. In their work, Gräbner et al. [12] proposed a classification system of the reviews of hotel customers employing sentiment analysis. Given a corpus, they designed a process to collect words that are related semantically and developed a domain-specific lexicon. This lexicon served as the key resource to develop a classifier for the reviews. The authors claimed to achieve a classification accuracy of 90%. Liu [14] presented a minutely detailed work on sentiment analysis and opinion mining. The author gave an in-deep introduction and presented a thorough survey of the available literature and the latest developments in the realm. This work presents an excellent qualitative and quantitative analysis of opinions and sentiments and stands as a distinguished literary resource for practical applications. The author endeavored to develop a common framework to bring different research works under a single roof and discussed the integral constituents of the subject like document-level sentiment classification, sentence-level subjectivity and sentiment classification, aspect-based sentiment analysis, sentiment lexicon generation, opinion summarization, and opinion spam detection. Bagheri et al. [3] proposed an unsupervised and domain-and language-independent model for analyzing online customers' reviews for sentiment analysis and opinion mining. Their generalized model was equipped with a set of heuristic rules to detect the impact of opinion word/multi-word. They presented a novel bootstrapping algorithm and proposed a metric to detect and score for implicit and explicit aspects of reviews. They claimed that their model can be used in a practical environment where high precision is required. Medhat et al. [17] gave a detailed analysis of sentiment analysis algorithms and their applications. Their work can be considered as the state of the art in the domain. The authors categorized a large number of research articles according to their participation in sentiment analysis techniques for real-world applications. They suggested that further research is needed to enhance Sentiment Classification (SC) and Feature Selection (FS) algorithms. They commented that the *Naïve Bayes* and *Support Vector Machine (SVM)* serve as the base or reference techniques for comparing novel SC algorithms. Fang and Zhan [10] proposed a general process for sentiment polarity categorization thereby giving detailed descriptions. They studied online product reviews from *Amazon.com* over the following major categories: beauty, books, electronics, and home appliances. Each review includes rating and review text among other data. The authors used a part of speech (POS) tagger at the preprocessing step and then computed sentiment score. They used the F1 measure to evaluate the performance of their proposed classification process and reported that the *SVM* model and the *Naïve Bayes* model performed almost the same. Mozetič et al. [19] exploited a big set of tweets from different languages. The tweets were labeled manually and they exploited them as training data. They proposed automatic classification models and reported that the performances of the top classification models are not statistically different. The authors concluded that it would be good to give more attention to the accuracy of the training data than the genre of the model being employed. They found that on applying to the three-class sentiment classification problem, there is no correlation between the accuracy and performance of the classification models. From the literature available on sentiment analysis, this work on human annotation is very unique. Saad and Saberi [22] presented a survey of sentiment analysis and opinion mining techniques, their applications, and challenges. The authors classified such techniques into three groups: machine learning approach, lexicon-based approach, and combination method. They collected data from blogs and forums, reviews, news articles, and social networks (*Twitter* and *Facebook*). They concluded that the unstructured data is a big hurdle in sentiment analysis and stated that algorithms of sentiment classification and opinion mining need further research for improvement. Ghag and Shah [11] mentioned that the *bag-of-words* is a popular tool of sentiment analysis. The authors classified the sentences extracted from the sentiments by reviewing their syntactic and semantic structures. They proposed some metrics like relative frequency, term frequency, and inverse document frequency to improve accuracy. They used text preprocessing techniques and claimed to achieve 77.2% classification accuracy.

Employing automatic text clustering and manual qualitative coding, Mäntylä et al. [16] analyzed around seven thousand research articles from *Scopus* and *Google scholar* and presented a computer-assisted literature review for sentiment analysis. They highlighted a very interesting fact that automatic sentiment analysis had been possible only with the availability of online subjective texts and therefore, 99% of the research work in this domain took place after 2004. According to the authors, computer-based sentiment analysis started by analyzing product reviews available over the web, and it is now being applied over a wide range of domains like social media texts (*Twitter*, *Facebook*, etc.), stock markets, elections, disasters, medicine, and cyberbullying. They stated that sentiment analysis involves a multitude of data sources like tweets, comments, chats, emoticons etc. Alsaeedi and Khan [2] investigated applications and results of various sentiment analysis techniques over *Twitter* data. They explored machine learning, lexicon-based approaches, ensemble approaches, and hybrid approaches. They reported the following conclusions of their research work: when multiple features were taken, machine learning techniques resulted in the greatest precision; lexicon-based techniques performed good but they require manual efforts to create the archive, and the ensemble and hybrid-based algorithms performed better than supervised machine learning algorithms. Tyagi and Tripathi [26] also collected *Twitter* data and performed sentiment analysis. The authors extracted the features through the N-gram modeling technique and exploited the *K-Nearest Neighbor* algorithm to categorize sentiments into positive, negative, and neutral. Bhagat et al. [6] studied online product reviews, general tweets in *Twitter*, and movie reviews and carried out sentiment analysis of text messages using supervised machine learning techniques. They preprocessed the messages and applied *Naïve Bayes*, *Decision Tree*, and *Support Vector Machine (SVM)* techniques for their research. They proposed a three-tier framework: the first layer is the initialization layer for data collection and message preprocessing, the second layer is the learning layer which splits preprocessed data into training and test datasets and develops three machine learning models, the final layer evaluates the performance of the models based on precision, recall, F1-measure, etc. The authors concluded that the *Decision Tree* and *SVM* can be considered as good classifiers with lower mean square error.

3 Data Collection and Methodology

As we all know, *Amazon* is one of the leading E-commerce websites, where a large number of users' reviews can be found. After purchasing the products, customers can post their reviews directly on the *Amazon* review portal. With such a massive amount of customers' reviews, this provides an opportunity to study and investigate feedbacks of the customers about a specific product [8]. All such comments or feedbacks help the sellers and other potential customers comprehend product-related public opinions. In the present case study, we are taking reviews of *Amazon* customers for sentiment analysis. We propose a *three-tier model* that takes raw data from the *Amazon* portal as input and generates a comparative report over certain parameters. In the first phase, we fetch data from the portal, apply data preprocessing and cleaning techniques to repair and/or remove the dirty data. In the second phase, we apply *TF-IDF* and *Skip-Gram* models for statistical feature selection. This step filters

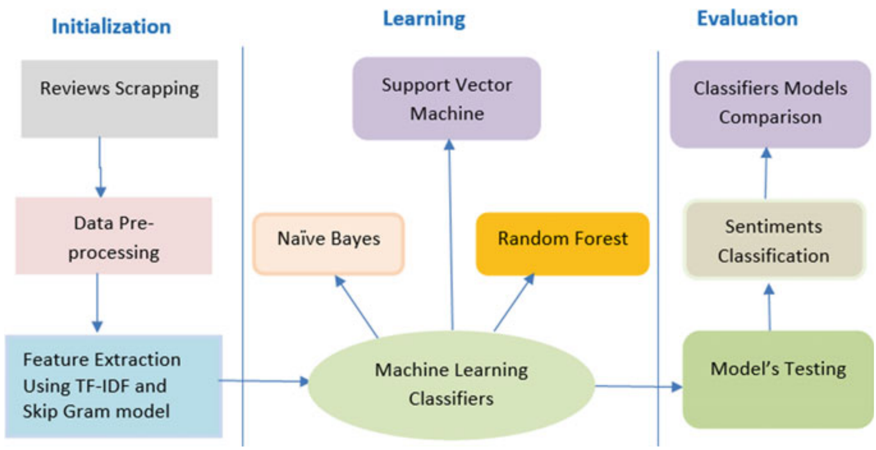


Fig. 1 Proposed model and its constituents

out those input variables which exhibit the strongest relationship with output variables. In the final phase, we apply machine learning (ML) algorithms say, *Random Forest*, *Naïve Bayes*, and *Support Vector Machine* analyzing the processed dataset and to classify the customers' reviews into the genres of positive, negative, and neutral. These algorithms are applied to the processed dataset to study the following performance parameters: accuracy, precision, recall, F measure, true positive, and false negative. Finally, our study compares the outputs of these three classifiers over the above-mentioned parameters. Figure 1 depicts the elements of our proposed model.

We hypothesize a four-fold methodology for our present research work: (a) data collection, (b) data preprocessing, (c) data representation, and (d) data classification. We now discuss the above steps in detail below.

3.1 Data Collection and Preprocessing

To conduct this case study, we gathered data from *the Amazon* web portal using an automated technique known as *Scrapping*. Scraping is a data extraction technique used for data collection from different websites. *Scrapy* is a free and open-source web-crawling framework which is written in *Python* and it is used for extracting data from websites. We applied the scraping process to extract *Amazon* reviews using *the Scrapy* library which permits the programmers to extract the data as per their requirements [18]. In our practical experiments, the scrapped data set consists of 300,000 mobile phone reviews from *the Amazon* review portal for various international brands. However, these reviews are in unstructured and unlabeled text form which requires pre-processing treatment. This is an essential step of the whole process as the accuracy of machine learning models depends on the quality of data we feed into them.

The scrapped data set used in our research had many missing or null values. We dealt with these issues by utilizing the *Imputation* technique, a widely used tool in the realms of machine learning and data mining. The basic principle of the technique is to replace each missing value of an attribute with the mean of the observed values of the attribute, known as Mean Imputation (MEI), or a nominal attribute with its most commonly observed value, known as Most Common Imputation (MCI). For each attribute f_i with missing values, the classifier C_i (. . .) takes as input the values of the other (*n* – *1*) attributes $\{f_j \mid j \neq i\}$ for an instance, and returns the value for f_i for this instance [4, 15, 24]. Other preprocessing treatments applied to this data before feeding the data to the machine learning models are spellings corrections; stop words removal; removal of special characters and punctuations from text data; removal of multiple spaces; removal of numeric digits from the review texts; removal of all URLs, hashtags, and E-mail addresses; upper to lower case conversion; contraction to expansion; substitution of any non-UTF-8 character by space; stemming; and removal of rare words. To improve the performance of the classifier's models, some of the irrelevant attributes (like reviews.dateSeen, reviews.sourceURLs, reviews.title, reviews.username, etc.) have been dropped after pre-processing.

Ultimately, after applying all of the above-mentioned preprocessing treatments, we receive accurate, useful, and clean text suitable for analysis and classification of sentiments. Table 2 presents the final extracted attributes and their description.

Every product rating is based on a 5-star scale ranged from 1-star to 5-star with no existence of a half-star or a quarter-star. Figure 2 depicted below shows the distribution of reviews based on *Amazon's* 1–5-star rating scales.

As shown in the above figure, the most frequent review rating in our dataset is 5 stars, with more than 30% share in the entire dataset. Figure 3 illustrated below shows the attributes which are of the numerical type and their distribution in the data set.

It is clear from the above figure that *reviews.numHelpful* is a valuable attribute in our dataset, so we kept only those instances in the dataset for which more than 75 people found the review helpful. On the other hand, in *reviews.rating* attribute, the distribution is skewed towards 5 stars rating. The last two attributes, *reviews.userCity* and *reviews.userProvince* have *NaN* values i.e., a numerical value that is undefined or not present. Therefore, we have dropped these attributes from our dataset. One important attribute that is used for product identification is *Amazon Standard Identification Number* (*ASIN*). Our dataset has 35 different products which possess unique *ASIN* values and are used for training our classifiers. After analyzing *ASINs* and product name attributes, we observed that there's a one to many relationships between the *ASINs* and the product names, i.e., a single *ASIN* is linked with one or more product names. Figure 4 shown below visualizes the individual *ASIN* and product reviews in a bar graph representation.

The above figure clearly shows that certain products have significantly more reviews than other products, which indicate a higher sale of those products. Based on this *ASIN* attribute frequency graph we can easily decide which products should be kept or dropped. Now, for better insight into the data or corpus, the *Wordcloud* visualization is an excellent tool in practice. The word that appears more prominent

Table 2 Features information of *Amazon* reviews dataset

S. No.	Attributes name	Description	Data type
1	id	This attribute represents a unique reviewer id number being assigned to each reviewer	Int64 (integer)
2	name	This attribute represents the product name	String
3	ASIN	This attribute stands for *Amazon Standard Identification Number*, a unique number assigned to each product	Object (integer and characters)
4	brand	It represents the brand name of the product	String
5	categories	It shows the category the product belongs to	String
6	manufacturer	This attribute represents the manufacturer of the product	String
7	reviews.date	This attribute expresses the date on which review posted on the *Amazon* portal	Float64 (floating point number)
8	reviews.doRecommend	This attribute shows whether a particular review is recommended by other reviewers or not	Boolean (yes/no)
9	reviews.numHelpful	The number of reviewers who consider a particular review useful	Int64 (integer)
10	reviews.userCity	This attribute represents reviewer city	String
11	reviews.userProvince	This attribute represents reviewer province or state	String
12	review rating	This attribute expresses the rating (1–5 stars) assigned to a particular product by a reviewer	Float64
13	review text	This attribute contains the actual review text posted by reviewers on the *Amazon* portal	Object (integer and characters)

Fig. 2 Distribution of *Amazon's* star-rating scores

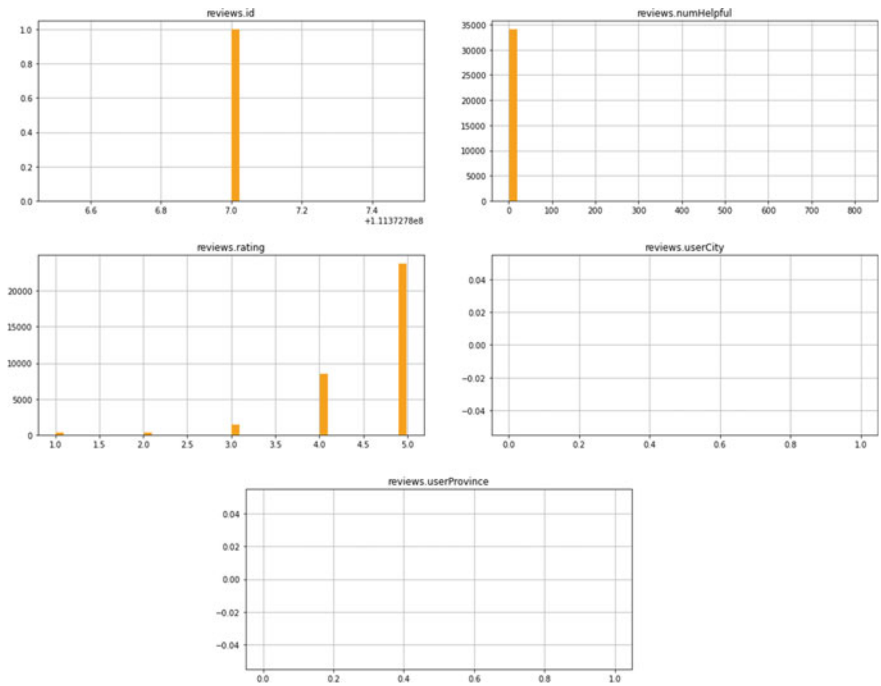


Fig. 3 Distribution of numerical data in the dataset

based on the frequency or importance in the text data is displayed with the bigger size in the *Wordcloud* visualization. In simple words, the word with larger size has more weight than the word with smaller size. After the completion of pre-processing of the dataset, we visualize words from the reviews' text using *Wordcloud* feature as shown in Fig. 5 below.

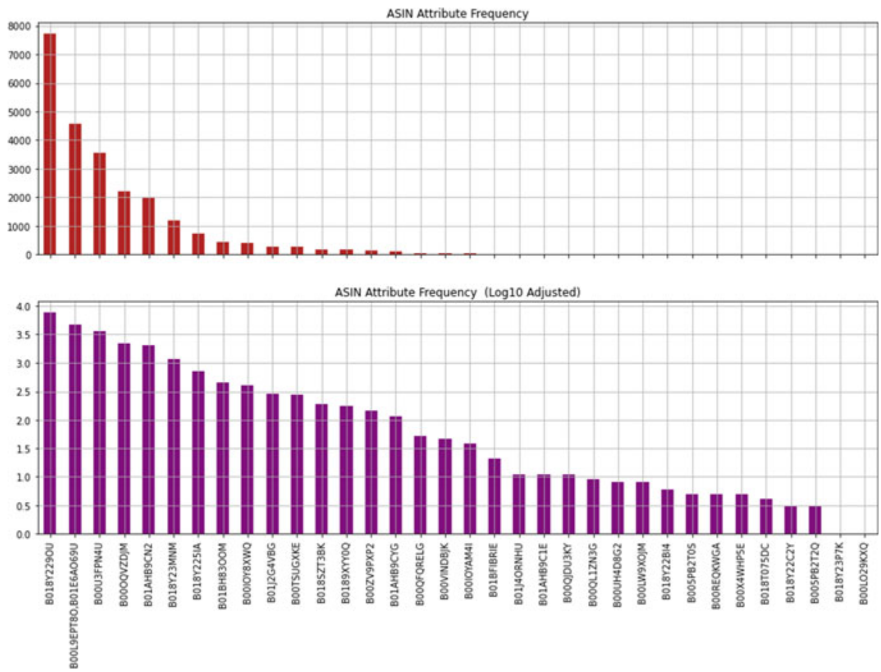


Fig. 4 Review ratings and ASIN frequencies

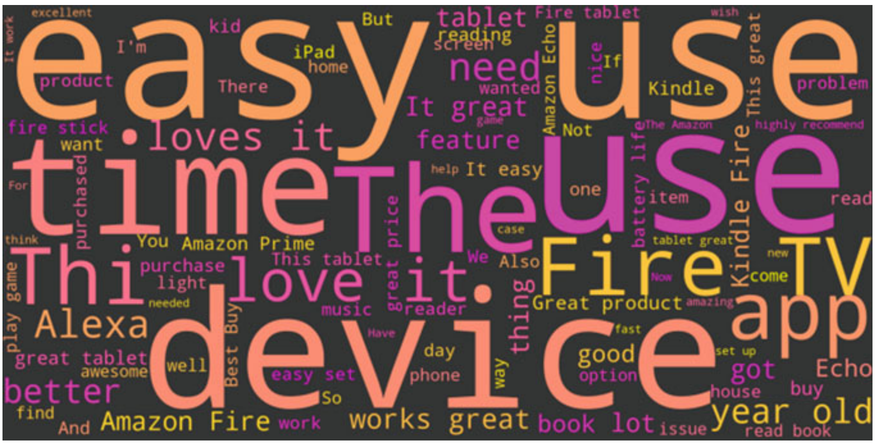


Fig. 5 *Wordcloud* of reviews' text

3.2 Data Representation

After the pre-processing of the unstructured text, the data representation is a vital step in sentiment classification. The extracted pre-processed reviews are mainly in text format but numerical representation in terms of metrics is needed to classify sentiments using the machine learning algorithms. Therefore, we have applied two different approaches to convert text data into some suitable form to be fed into the machine learning classifiers. The first approach is *word embedding* and second is the combination of *Term frequency and inverse document frequency (TF-IDF).* For word embedding, we applied the *Word2vec* model with *skip-gram* architecture. The *skip-gram* model predicts the source context words given a target word. It works as an unsupervised learning technique that is used to find the most suitable and related words for a given target word [13]. *Skip-gram* architecture provides more accurate and effective results when we have a corpus of bigger size, because, in the *skip-gram* approach, each context-center pair is considered as a new observation. The word vectors are adapted using Eq. (3.1), as given below:

$$w_{i,j}(k+1) = w_{i,j}(k) - s\frac{\partial J}{\partial w_{i,j}} \tag{3.1}$$

where $w_{i,j}(k)$ is word vector value in step k of the optimization process, j is our *optimization function* and s is the chosen step size. The *optimization function* is applied for selecting those words which can be represented using the Eq. (3.2) given below.

$$J = \sum_{i,j=1}^{V} f\left(X_{i,j}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{i,j}\right)^2 \tag{3.2}$$

where V is the number of word tuples with the non-zero co-appearance count, $X_{i,j}$ is the count of co-appearances, w_i is a word vector and $\tilde{w}_j$ is word vector's context, b_i and $\tilde{b}_j$ are biases (again every word has two of them: one for the word and other for the context) and function f is a weighing function. The *skip-gram* architecture is illustrated in Fig. 6 given below.

The *TF-IDF* algorithm is based on words' statistics for feature extraction and represents how important a word or a phrase in a corpus. *TF-IDF* assigns a unique score to each word using a hybrid statistical method, in terms of the product of *term frequency* (TF) with *inverse document frequency* (IDF). The TF denotes the total number of times a given term occurs in the dataset against the total number of all words in the document, and the IDF measures the amount of information word provides [23]. In our case study, TF assigns a score to most frequently occurring words in the mobile review dataset and IDF assigns weight to the least frequent words in the same dataset. The *TfidfVectorizer* from the *sklearn python library* is used to fit the vectorizer on the corpus of the review texts to extract features and the model will transform the text data into the *TF-IDF* representation. Using the *TF-IDF*

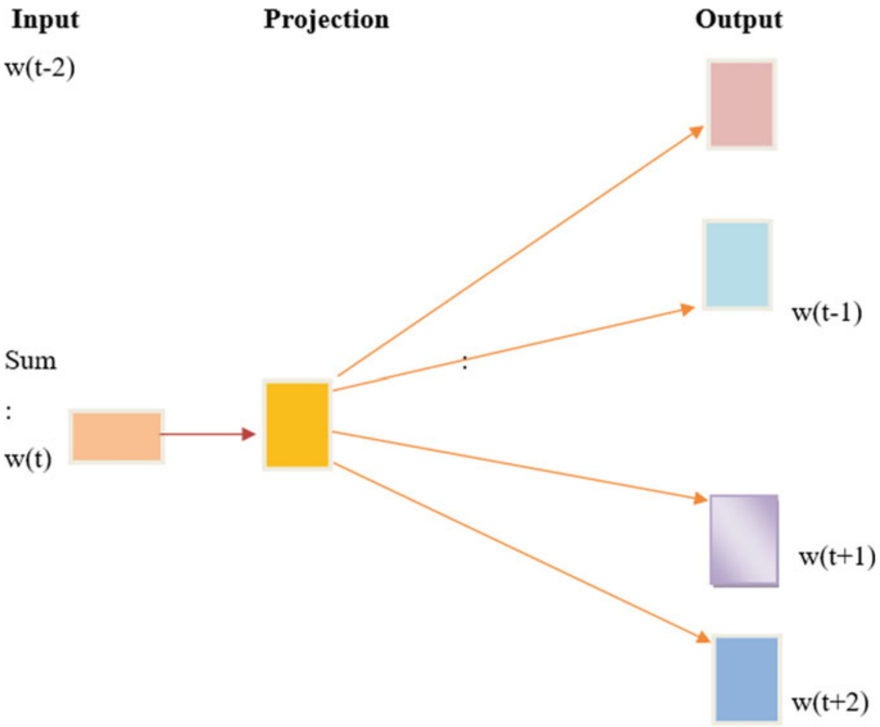


Fig. 6 Architecture of word to vector skip-gram model

approach in a normalized data format, each corpus word can be represented using the following equation:

$$W_{ij} = tf_{ij} \times \log\left(\frac{N}{df_i}\right) \tag{3.3}$$

where, tf_{ij} is the number of occurrences of the ith word in the jth review and df_i is the number of reviews containing i and N is the total number of reviews.

3.3 Classifications

Classification is a supervised machine learning process that generally focuses on predicting a qualitative response by recognizing and analyzing given data points. This case study is focused on sentiment analysis by classifying the reviews into 1–5 ranked scales. To carry out this task, we have applied three different supervised classifiers: *Naïve Bayes*, *Support Vector Machine*, and *Random Forest*. These classifiers are then evaluated to provide a comparative analysis of various parameters for classifying reviews into positive and negative genres. The literature suggests that reviews with ratings 4 and 5 should be categorized as positive reviews while reviews with rating 3 should be labeled as neutral and reviews with ratings 2 and 1 should be treated as negative reviews. Since here we are interested in analyzing only positive and negative reviews, so we have neglected neutral reviews form the data set. The working of the individual classifier is explained in the next sections.

3.3.1 Naïve Bayes Classifier

The *Naïve Bayes* (NB) classifier is a simple and robust probabilistic classifier algorithm that is based on the *Bayes* theorem. It assumes that attribute values are independent of each other given the class. This assumption is known as the *conditional independence assumption*. Therefore, applying changes in one feature does not affect other features of the class [7]. Let D be our *Amazon* review data set for training the model then each tuple in the dataset is defined with n attributes and it is represented by: $X = \{a_1, a_2, a_3, \ldots, a_n\}$. Let there be m classes represented by: $\{C_1, C_2, C_3 \ldots, C_m\}$. For a given tuple X, the classifier predicts that X belongs to the class having the highest posterior probability, conditioned on X. The *Naïve Bayes* classifier predicts that the tuple X belongs to the class C_i if and only if $P(C_i|X)$ is maximum among all i.e.:

$$P(C_i|X) > P\left(C_j|X\right) for 1 \le j \le m, i \ne j \tag{3.4}$$

Since we want to maximize $P(C_i|X)$, the class C_i for which $P(C_i|X)$ is maximized is called the *maximum posterior hypothesis*. According to the *Bayes* theorem,

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)} \tag{3.5}$$

If the attribute values are conditionally independent of one another (*Naïve Bayes* condition), then

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{3.6}$$

where x_k refers to the value of the attribute A_k for the tuple X. If A_k is a categorical attribute, then $P(x_k|C_i)$ is the number of tuples of class C_i in D having the value x_k for A_k. The classifier predicts the class label of X is in the class C_i if and only if,

$$P(X|C_i) \cdot P(C_i) > P(X|C_j) \cdot P\left(C_j\right) \tag{3.7}$$

3.3.2 Support Vector Machine

Support Vector Machines (SVM) is a supervised machine-learning based classification algorithm which widely deals with predictive and regression analysis. *SVM* algorithm aims to find a *hyperplane* in an N-Dimensional feature space that distinctly classifies the data points, while maximizing the marginal distance for the two classes (positive and negative) and minimizing the classification errors. The marginal distance for a class is the distance between the decision hyperplane and its nearest instance which is a member of that class [25]. The data points that lie closest to the decision surface (or hyperplane) are called support vectors and these points help us in building the *SVM* model. The loss function that helps in maximizing the margin is given below.

$$C(x, y(f(x)) = f(x) = \begin{cases} 0, & x < 0 \\ 1 - y * f(x) & else \end{cases} \tag{3.8}$$

$$C(x, y(f(x)) = (1 - y * f(x))+ \tag{3.9}$$

The equation of the line in 2D space is $y = a + bx$. By renaming x with x_1 and y with x_2, the equation will change to $ax_1 - x_2 + b = 0$. If we specify $X = (x_1, x_2)$ and $w = (a, -1)$, we get,

$$F(x) = w \cdot x + b \ where \ w, x \in R^n \ and \ b \in R \tag{3.10}$$

The above Eq. (3.10) is called the equation of the hyperplane, which linearly separate the data.

The hypothesis function h in SVM classifier can be defined as:

$$h(x) = \begin{cases} +1 \ if w.x + bx \geq 0 \\ -1 \ \ if w.x + b < 0 \end{cases} \tag{3.11}$$

The point above or on the hyperplane will be classified as class $+1$, and the point below the hyperplane will be classified as class -1. SVM classifier amounts to minimizing an expression of the form given below:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{max(0, 1 - yi(w.x_i - b))}\right]$$

The reason to choose this classifier in the present research is its robustness. It provides an optimal margin gap to separate hyperplanes and it gives a well-defined boundary for easy classification.

3.3.3 Random Forest

Random Forest (RF) machine learning algorithm is also known as a tree-based ensemble learning, which creates a forest of many decision trees. RF ensures that the behavior of each decision tree produced is not too correlated with the behavior of any other decision tree in the model. This final prediction can simply be the mean of all the observed predictions [9]. Therefore, the different decision trees obtained by using the RF algorithm are trained using different parts of the training dataset, which is the reason behind its unbiased nature and superior prediction accuracy.

4 Experimental Results and Discussions

For conducting the practical implementation of this case study, we used *Jupyter Notebook* with Python version 3.8. Various Python libraries have been used for data pre-processing and visual representation such as *pandas, numpy, scrapy, matplotlib, seaborn, spacy,* etc. For training and testing of machine learning classifiers, the corpus is divided into two subsets with a train-test split of 75–25% respectively. To evaluate the performances of the classifiers, the main parametric metrics employed in this research are *accuracy, precision, recall, F-measure, true positive* and *false negative*. In classification problems, precision (also called positive predictive value) is the fraction of relevant instances that are retrieved, while recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Accuracy of the model can be defined as the ratio of the correctly labelled attributes to the whole pool of variables. F-measure is a weighted average of precision and recall. As *ASIN* and review ratings are two important attributes available in our dataset, therefore we explored their relationship visualized in Fig. 7 below.

The above figure clearly reveals that the most frequently reviewed products in our case study have their average review ratings above 4.5. On the other hand, *ASINs* with lower frequencies in the bar graph have their corresponding average review ratings below 3. For analyzing the classification performance of machine learning models, we applied *Naïve Bayes*, *Support Vector Machine* and, *Random Forest* algorithms to the pre-processed dataset. The performance evaluation results of machine learning classifiers using *Skip-gram* and *TF-IDF* feature extraction techniques are shown in Table 3 and Table 4 respectively.

The results of the Table 3 above, clearly show that the *Random Forest* classifier achieves maximum accuracy in skip-gram model.

Table 4 reveals that *TF-IDF* significantly improves the accuracy along with other important parameters of *Naïve Bayes* and *Support Vector Machine* classifiers but does not perform well enough with *Random Forest*. After comparing the results of Table 3 and Table 4, it is clear that the *TF-IDF* approach improves the accuracies of *Naïve Bayes* and *Support Vector Machine* classifiers by 6% and 3.9% respectively but deteriorates the accuracy of *Random Forest* by a margin of 0.8%. Figure 8 compares

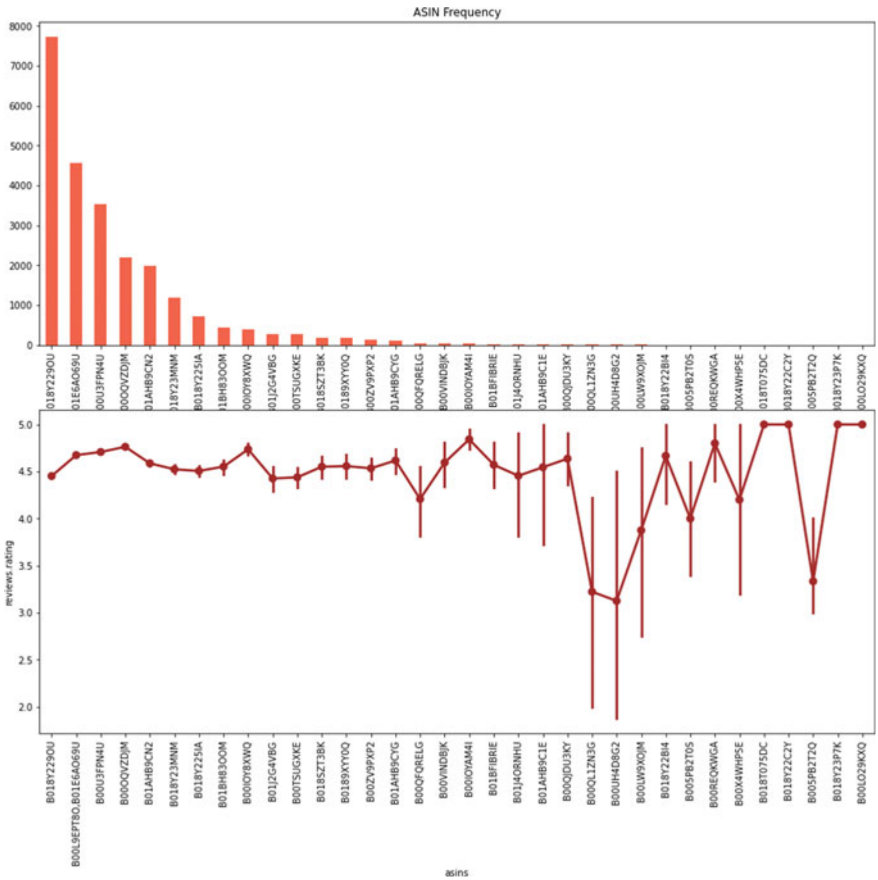


Fig. 7 Relationship between *ASIN* and review ratings

Table 3 Classifiers comparison using Skip-gram feature extraction method

Classifier	Accuracy	Precision	Recall	F-measure
Naïve Bayes	89.3	0.893	0.892	0.892
Random Forest	95.2	0.952	0.95	0.951
Support Vector Machine	93.3	0.916	0.93	0.927

Table 4 Classifiers comparison using *TF-IDF* feature extraction method

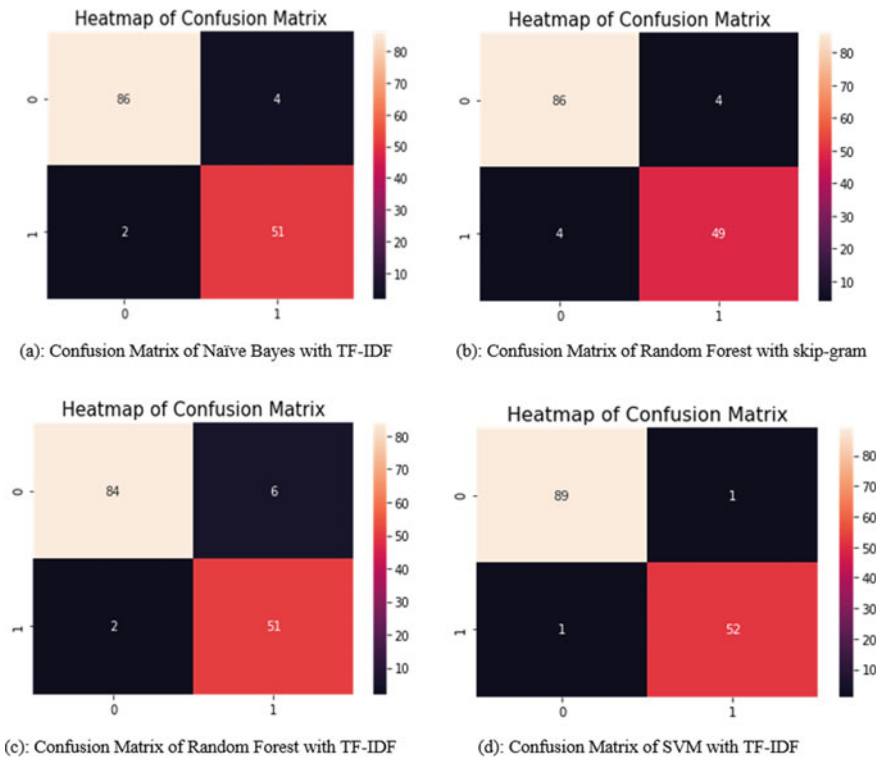
Classifier	Accuracy	Precision	Recall	F-measure
Naïve Bayes	95.33	0.953	0.952	0.953
Random Forest	94.63	0.946	0.941	0.946
Support Vector Machine	98.61	0.991	0.99	0.990

Fig. 8 Accuracies comparison of *Skip-gram* and *TF-IDF* techniques

the classification accuracies of individual classifiers against the *Skip-gram* and *TF-IDF* feature extraction techniques.

As shown in the above figure, it is evident that in the case of *Naïve Bayes* classifiers, the classification accuracy obtained using *TF-IDF* is better than the value obtained using the *Skip-gram* technique. The *Naive Bayes* algorithm follows a probabilistic approach, where the attributes are independent of each other. Therefore, when the analysis is performed using a single word (unigram) and double word (bigram), the accuracy value obtained with *TF-IDF* is comparatively better than that obtained using *Skip-gram*. Similarly, it is clear that in the case of *Random Forest*, the classification accuracy value obtained using the *Skip-gram* technique is a little better than the value obtained using *TF-IDF*. As we know, *Random Forest* is an ensemble tree-based classifier and it aggregates the output obtained from different decision trees, the *Skip-gram* model which can predict the source context words given a target word gives better results. In the case of *Support Vector Machine* classifiers, the classification accuracy attained using *TF-IDF* is better than that obtained using the *Skip-gram* approach. *Support vector machine* is a non-probabilistic linear classifier and the trained classifier is used to find hyperplane for dataset separation, the *TF-IDF* which analyses the corpus word by word gives better results as compared to the *Skip-gram* model. Figure 9 shown below presents the *Heatmaps* of confusion matrices obtained.

The above figure depicts the four best confusion matrices obtained from various classifiers, which is a summary of the prediction results in our classification problem. The part (a) above, shows the trained model predicts *True Negative* of 86%, *True Positive* of 51%, *False Positive* of 4% and *False Negative* with 2%. Therefore, 86% and 51% are the correct predictions and 4% and 2% are incorrect predictions. We



(a): Confusion Matrix of Naïve Bayes with TF-IDF
(b): Confusion Matrix of Random Forest with skip-gram
(c): Confusion Matrix of Random Forest with TF-IDF
(d): Confusion Matrix of SVM with TF-IDF

Fig. 9 Heatmaps of the confusion matrices

can see that we have a good ratio of correct predictions. Similarly, we can interpret the confusion matrices of other classifiers. Note that the SVM with TF-IDF shows maximum accuracy, as shown in part (d) of Fig. 9.

5 Conclusions

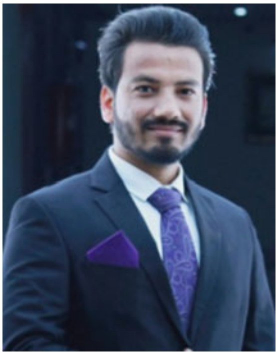
Analysis of sentiments is crucial for any online retail business enterprise to understand the opinions and feedbacks of its customers. This case study analyses the sentimental polarities of the scrapped user-reviews of *Amazon* customers through machine learning classifiers. The dataset used in the present chapter is collected from the *Amazon* review portal using the well-known *scrappy* library available in Python. We scrapped 300,000 mobile phone reviews from *Amazon* review portal for various international brands. This unstructured dataset had to be preprocessed first to convert it into a legitimate form so that machine learning classifiers can process it smoothly. The null and missing values were dealt with *imputation* technique and several preprocessing techniques like stop word removal, spelling corrections, stemming, special character handling, etc. were also exercised. Now, this preprocessed dataset which was in text form needed to be converted into numerical scores before submitting it into machine learning models. We employed two techniques for this preliminary step: *Skip-gram* and *TF-IDF*. After the above treatments, we put the processed dataset to three different classifiers: *Naïve Bayes*, *Support Vector Machine*, and *Random Forest*. The above-mentioned machine learning classifiers are then evaluated over some standard parameters say, accuracy, precision, recall, F-measure, true positive and false negative. The empirical results found to be very satisfactory.

The present sentiment analysis case study of the *Amazon* reviews can be considered a kind of novel work where various machine learning classifiers have been compared against two different feature engineering techniques. Empirical results reveal that all models are able to classify the user reviews into negative and positive classes with relatively high accuracy and precision. Calculated results exhibit that the *Support Vector Machine* model achieved the highest classification accuracy (98.61%) with *TF-IDF* feature extraction method. Next, the *Naïve Bayes* model with *TF-IDF* achieved the classification accuracy of 95.33%. And, the *Random Forest* model with the *Skip-gram* technique acquired 95.2% accuracy at the third position.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Proceedings of the Workshop on Languages in Social Media (LSM 2011), pp. 30–38. Association for Computational Linguistics (2011)
2. Alsaeedi, A., Khan, M.Z.: A study on sentiment analysis techniques of Twitter data. Int. J. Adv. Comput. Sci. Appl. **10**(2), 361–374 (2019)
3. Bagheri, A., Saraee, M., Jong, F.D.: Care more about customers: unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowl.-Based Syst. **52**, 201–213 (2013)
4. Bai, B.M., Mangathayaru, N., Rani, B.P.: An approach to find missing values in medical datasets. In: Proceedings of the International Conference on Engineering & MIS 2015, Istanbul Turkey, Article No.: 70, pp. 1–7 (2015). https://doi.org/10.1145/2832987.2832991
5. Barbosa, L., Feng, J.: Robust sentiment detection on Twitter from biased and noisy data. In: Conference: COLING 2010, 23rd International Conference on Computational Linguistics, Beijing, China, Poster Volume, pp. 36–44 (2010)
6. Bhagat, A., Sharma, A., Chettri, S.K.: Machine learning based sentiment analysis for text messages. Int. J. Comput. Technol. **7**(6), 103–109 (2020)
7. Celin, S., Vasanth, K.: ECG signal classification using various machine learning techniques. J. Med. Syst. **42**(12), 1–11 (2018). https://doi.org/10.1007/s10916-018-1083-6
8. Chaffey, D.: Amazon's business strategy, revenue model and culture of metrics: a history. Digital Marketing case studies (2020). https://www.smartinsights.com/digital-marketing-strategy/online-business-revenue-models/amazon-case-study/. Accessed 25 July 2020
9. Dimitriadis, S.I., Liparas, D.: Alzheimer's disease neuroimaging initiative. How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from alzheimer's disease neuroimaging initiative (ADNI) database. Neural Regen. Res. **13**, 962–970 (2018). http://www.nrronline.org/text.asp?2018/13/6/962/233433

10. Fang, X., Zhan, J.: Sentiment analysis using product review data. J. Big Data **2**(5), 1–14 (2015)
11. Ghag, K.V., Shah, K.: Conceptual sentiment analysis model. Int. J. Electr. Comput. Eng. **8**(4), 2358–2366 (2018)
12. Gräbner, D., Zanker, M., Fliedl, G., Fuchs, M.: Classification of customer reviews based on sentiment analysis. In: Fuchs, M., Ricci, F., Cantoni, L. (eds.) Information and Communication Technologies in Tourism 2012, pp. 460–470. Springer, Vienna (2012). ISBN 978-3-7091-1141-3
13. Krishna, P.P., Sharada, A.: Word embeddings—skip gram model. In: Gunjan, V., Garcia Diaz, V., Cardona, M., Solanki, V., Sunitha, K. (eds.) ICICCT 2019—System Reliability, Quality Control, Safety, Maintenance and Management. ICICCT 2019. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-8461-5_15
14. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan&Claypoo, SanRafael, CA (2012)
15. Madhu, G., Rajinikanth, T.V.: A novel index measure imputation algorithm for missing data values: a machine learning approach. In: IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, pp. 1–7 (2012). https://doi.org/10.1109/iccic.2012.6510198
16. Mäntylä, M.V., Graziotin, D., Kuutila, M.: The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. Comput. Sci. Rev. **27**, 16–32 (2018). ISSN 1574-0137
17. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
18. Mitchell, R.: Web Scraping with Python, 2nd ed. O'Reilly Media, Inc. (2018). ISBN: 9781491985571
19. Mozetič, I., Grčar, M., Smailović, J.: Multilingual Twitter sentiment classification: the role of human annotators. PLoS ONE **11**(5), (2016). https://doi.org/10.1371/journal.pone.0155036
20. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
21. Prabowo, R., Thelwall, M.: Sentiment analysis: a combined approach. J. Inform. **3**(2), 143–157 (2009)
22. Saad, S., Saberi, B.: Sentiment analysis or opinion mining: a review. Int. J. Adv. Sci. Eng. Inf. Technol. **7**(5), 1660–1666 (2017)
23. Sidorov, G.: Vector space model for texts and the tf-idf measure. In: Syntactic n-grams in Computational Linguistics. Springer Briefs in Computer Science. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14771-6_3
24. Su, X., Greiner, R., Khoshgoftaar, T.M., Napolitano, A.: Using classifier-based nominal imputation to improve machine learning. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) Advances in Knowledge Discovery and Data Mining. PAKDD 2011. Lecture Notes in Computer Science, vol. 6634. Springer, Berlin, Heidelberg (2011)
25. Sweilam, N.H., Tharwat, A.A., Moniemc, N.K.A.: Support vector machine for diagnosis cancer disease: a comparative study. Egypt. Inform. J. **11**(2), 81–92 (2010)
26. Tyagi, P., Tripathi, R.C.: A review towards the sentiment analysis techniques for the analysis of Twitter data. In: Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) (2019)



Er. Koushal Kumar has 8 years of teaching experience at both graduate and postgraduate levels. He attended several professional development seminars, national and international conferences, Workshops, and faculty development programmes. He also cleared India's prestigious UGC-NET examination in 2013. He has been actively performing reviewer duties in an IEEE journal (Internet of things) since 2017. He has more than 130 citations and 12500 reads on world most reputed researcher platform Research Gate. He has guided many dissertations at the master's level. He is a keen researcher and has published many research publications in international journals and conferences. His research interests include Machine Learning, Data science, Internet of Things, Fog computing and Deep Learning. He is the author of a book titled "*Rules extraction from trained neural networks using decision trees*."

Dr. Bhagwati Prasad Pande has experience of more than 8 years in teaching undergraduate and postgraduate courses of Information Technology and Computer Science. He holds Ph.D. degree in IT and his research interests include Information Retrieval, NLP, Optimization Techniques and Emoticons. He also cleared India's prestigious UGC-NET examination in 2013. He published many research articles in reputed publication houses like John Wiley & Sons, Oxford University Press and Glottotheory. He recently registered copyright in *The United States Copyright Office*. He also served a few journals as a reviewer. He has guided a number of dissertations at the master's level and industrial projects at the undergraduate level.

Pattern Mining—FTISPAM Using Hybrid Genetic Algorithm

L. Mary Gladence, S. Shanmuga Priya, A. Shane Sam, Gladis Pushparathi, and E. Brumancia

Abstract An innovative approach is elegantly launched to effectively identify the medical behavioural changes of the patients. With this end in view, the sequential change patterns are extracted at two diverse time intervals, with the help of the fuzzy time interval sequential change pattern mining employing the HGA technique. However, the pattern mining at two diverse time intervals is likely to yield further superfluous data. With an eye on averting the generation of the corresponding superfluous data, an optimized method such as the hybrid genetic algorithm (HGA) based fuzzy time interval sequential pattern mining is envisaged for the purpose of attaining the patterns. The sequential pattern detection algorithm effectively segments the located change patterns into four diverse types such as the perished patterns, added patterns, unexpected changes, and the emerging patterns. When the pattern categorization comes to an end, the changed patterns are harmonized by means of the Similarity Computation Index (SCI) values. At last, the significant patterns are estimated and employed to categorize the change in the conduct of the patient. The imaginative system is performed in the working stage of the MATLAB and its execution is surveyed and appeared differently in relation to that of the advanced strategy like the genetic algorithm.

Keywords Pattern mining · Similarity computation index · E-Emerging patterns · E-Unexpected pattern · E-Added/Perished pattern

L. Mary Gladence (✉) · E. Brumancia
Sathyabama Institute of Science and Technology, Chennai, India

S. Shanmuga Priya
M.I.E.T Engineering College, Trichy, India

A. Shane Sam
St. Joseph's College of Engineering, Chennai, India

G. Pushparathi
Velammal Institute of Technology, Chennai, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_16

1 Introduction

The data mining has surfaced as an innovative approach devoted for the effective tackling of titanic data to discern the concealed patterns and linkages which go a long way in performing prudent decision making [1]. In fact, the data mining technique is well-equipped with the requisite skills of extracting the frequent patterns, locate the relationships, and effectively execute the categorization and forecast [2]. Consequently, the data mining shows its mettle in efficiently extracting the enthusing patterns from the databases leading to the incredible mitigation of the severe hurdles in the acquirement of knowledge in setting up smart systems [3]. The cardinal problem of data mining during the last few years has been invested in effectively extracting the useful sequence patterns from the given set [4].

In this regard, sequential pattern mining (SPAM) invites attention as one of the most appealing investigation topics in the realm of the data mining [5]. It is worth mentioning that the data mining approaches have been afforded a red-carpet welcome in many an application horizon thanks to its unique proficiency in the successful extraction of the sequential patterns from data [6]. In fact, the sequential pattern mining locates the frequent sub-sequences as the patterns in a sequence database. A sequence database, in turn, is entrusted with the task of storing a multitude of records, where all the records flow as the sequences of well-orchestrated events, with or without tangible concepts of duration [7]. In this regard, the mining of the sequential patterns, which is intended to locate the temporal relationships among the item-sets in the sequence database, has become an essential problem in the domain of the data mining [8].

Normally, the mining task has metamorphosed into a Herculean Task often eating up an incredible duration of time as the patterns can be generated by various transformations of item sets, which in turn, are created by any permutation of probable items in the database [9]. Of late, the mining of sequential patterns in the data streams has created a whirlpool of enthusiasm and animated activity among the intriguing investigators [10]. The sequential pattern mining, which represents one of the most vital sophisticated technologies, is competent to extract the patterns which occur with a higher frequency in relation to user-defined minimum support while preserving their item incidence order (Hirate and Yamana 2010).

A few of the most well-known techniques focused on the sequential pattern mining encompass the Apriori All, GSP (Generalized Sequential Patterns) and the Prefix Span [11]. The GSP and Apriori All represent the Apriori-like approaches taking a cue from the breadth-first investigation, whereas the Prefix Span technique is invariably dependent on the depth-first exploration [12]. There is also a feast of parallel techniques like the SPADE and the SPAM which are extensively employed in the ongoing investigations. The sequential pattern mining faithfully discharges the assignment of locating the sequential patterns which habitually happen in time sequence or specific order [10]. Thus, it is very easy to predict the upcoming scenarios, by means of a deft appraisal of the variation in the state of the sequences [13].

Consequently, the physicians are capable of identifying the fruition course of ailments with regard to the medical history of the subjects with the intention of averting and healing the diseases with no loss of time [14]. The major motive of the Sequential Pattern Mining technique is devoted to the determination of the frequent sequences in the specified database (Gladence 2016). It has, over the years, emerged as a significant data mining issue with extensive applications, like the web log mining, DNA mining, and so on. It represents a daunting challenge in as much it is essential to scrutinize the combinatorial explosive number of potential subsequent patterns [15].

One of the striking features of the Sequential pattern Mining is its innate skills in the effective identification of the frequent sub-sequences in a dataset [16]. Further, the sequential pattern mining is extensively applied in many an application. It has a vital part to play in the DNA Sequence patterns which are highly fruitful in the domain of the Medical applications [17]. It has also metamorphosed as an indispensable data mining function with several applications, encompassing market and customer appraisal, web log evaluation, pattern identification in the protein sequences, and extracting the XML query access patterns for the purpose of caching [18].

Sequential Pattern Mining is personally connected to hypothesis of the Data Mining. The area of the consecutive examples, which is an improvement over the successive thing set finding of affiliation rule mining, has developed as an overwhelming test thanks it's amazing unpredictability. The literature is flooded with a number of diverse algorithms and methods which are employed in the task of the sequential pattern mining. One of the out-of-the-ordinary modern sophisticated methods is the Mine Fuzz Change model, which is envisioned for the Sequential Pattern Mining process by employing the SCI (Similarity Computation Index) which effectively carries out the pattern classification process. But the captioned model is plagued by certain deficiencies in the SCI evaluation. In view of the fact that, the SCI value is estimated by means of the raw data which is gathered from diverse time intervals, it leads to the time intricacy in the pattern mining task. Further, there is a dearth of works related to the mining of the change in the patient behaviour in the fuzzy time-interval sequential patterns employing the Hybrid Genetic technique. When the entire deficiencies illustrated in the literary works are efficiently overwhelmed, the mining accuracy is augmented together with the superlative efficiency. However, the absence of effective solutions for the deficiencies has inspired us to carry out the investigation work in the area concerned.

2 Fuzzy Time-Interval Sequential Patterns Using Hybrid Genetic Algorithm

Sequential Pattern Mining is personally connected to hypothesis of the Data Mining. The area of the consecutive examples, which is an improvement over the successive thing set finding of affiliation rule mining, has developed as an overwhelming test thanks its amazing unpredictability [19]. The literature is flooded with a number of diverse algorithms and methods which are employed in the task of the sequential pattern mining. One of the out-of-the-ordinary modern sophisticated methods is the Mine Fuzz Change model, which is envisioned for the Sequential Pattern Mining process by employing the SCI (Similarity Computation Index) which effectively carries out the pattern classification process [20]. But the captioned model is plagued by certain deficiencies in the SCI evaluation. In view of the fact that, the SCI value is estimated by means of the raw data which is gathered from diverse time intervals, it leads to the time intricacy in the pattern mining task [21]. Further, there is a dearth of works related to the mining of the change in the patient behavior in the fuzzy time-interval sequential patterns employing the Hybrid Genetic technique [22]. When the entire deficiencies illustrated in the literary works are efficiently overwhelmed, the mining accuracy is augmented together with the superlative efficiency. However, the absence of effective solutions for the deficiencies has inspired us to carry out the investigation work in the area concerned [23, 24].

3 Overview of the FTISPAM-HGA

Here, an optimized FTI technique is envisaged for extracting the significant change patterns from the database which are gathered from two diverse time periods [25]. The new-fangled approach shown in “Fig. 1” flows through the following four phases such as (i) the fuzzy time interval sequential pattern mining employing the HGA algorithm (ii) the Patterns matching employing the SCI (iii) Patterns classification based on SCI value and (iv) the Significant pattern evaluation. In the initial phase, the patterns are extracted at two time intervals and are furnished to the pattern matching procedure employing the Similarity Computation Index (SCI), where the similarity is located for the patterns extorted at two diverse time instants. In the third phase, the patterns are effectively categorized. The patterns are sorted as the Emerging patterns, unexpected changes and added/perished patterns dependent on the SCI esteems which can be utilized to order the adjustment in lead of the patient [26].

4 Fuzzy Time Interval Sequential Pattern Mining

The input of pattern mining consists of a sequence database D_{seq}, a set $I_{itemset} = \{i_1, i_2, \ldots i_n\}$ of items and a set $T = \{I_1, I_2, \ldots I_m\}$ of time intervals, where T an absolute is and non-overlap segment of the time domain. A sequence $A = (a_1, y_1, a_2, y_2 \ldots a_{k-1}, y_{k-1}, a_k)$ is a time-interval sequence if $a_i \in I$ and $y_i \in T$ for $1 \leq i \leq k-1$ and $a_k \in I$. The outcome consist of all time-interval sequences which happen habitually in D_{seq}.

Two databases such as the D^{t_o} and D^t at diverse time intervals t_0 and t are employed to efficiently carry out the change pattern mining task, where the one length frequent

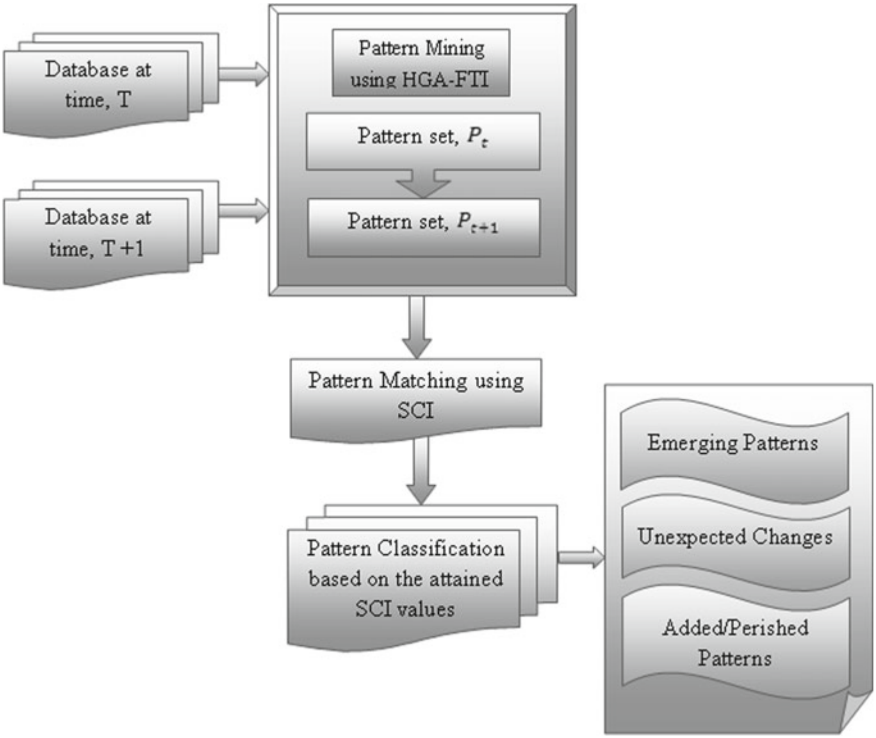


Fig. 1 Pattern classification with HGA

patterns are extracted in accordance with its support values. The support value, in turn, is defined as the number of tuples in a sequence database D_{seq} containing sequence seq called the support of seq, denoted as $s(seq)$. The patterns are mined which are greater support value than the offered user defined minimum support threshold value β and called as the frequent one length patterns. It is followed by extraction of the residual patterns by means of the optimization method viz the Hybrid Genetic Algorithm (HGA). As the unrefined data chosen at diverse time intervals contain greater redundant values resulting in the incredible intricacy of the task, it has to be effectively tackled by means of the further optimized approach such as the HGA [27]. In the earlier investigations, the genetic algorithm is employed as the optimization tool, though it is plagued by several deficiencies as shown below.

- It consumes a greater duration of time.
- It is likely to originate grave faults.
- There is a greater possibility of the optimal parameter values undergoing variation in the course of the process.

5 Hybrid Genetic Algorithm

With the intention of overwhelming the deficiencies of the genetic algorithm, the Hybrid Genetic Algorithm is effectively employed, which uses the versatile possibilities of crossover and mutation to accomplish the twin destinations of safeguarding assorted variety in the populace and keeping up the meeting aptitudes of the GA. In the Adaptive Genetic Algorithm (AGA), the probabilities of crossover and mutation changed in accordance with the fitness function values of the solutions.

Steps are given below

- Self-assertively create an underlying populace of people by methods for an emblematic portrayal procedure. Gauge the wellness of the people by assessing the bond vitality of the applicant arrangements portrayed by them.
- Pick two or three people from the present populace by methods for a customary roulette wheel choice administrator.
- The picked people are recombined to deliver two or three posterity by methods for the in part coordinated crossover as per the crossover rates. Gauge the wellness of the two posterity by assessing the bond vitality of the competitor arrangements portrayed by them.
- Place on record the level of development, or the level of hardship, on their wellness esteems because of the crossover function.
- Process the regular emblematic mutation of the two posterity and gauge their wellness by ascertaining the bond vitality of the competitor arrangements described by them.
- Place on record the level of advance, or the level of hardship, of their wellness esteems by virtue of the mutation function.
- Allocate the resultant people into another populace pool. On the off chance that the populace measure isn't achieved, return to step 2.
- Optimize the crossover and mutation rates according to the particular standards.

6 Fuzzy Time Interval Sequential Pattern Mining Using HGA Algorithm

The utilization of the optimization technique in the pattern mining goes a long way in scaling down the consumption of time in addition to being fruitful in the task of achieving most advantageous data from the database. The modus operandi of the HGA based pattern mining procedure is effectively exhibited in "Fig. 2".

Consequently, the patterns are mined during the time interval t_o, characterized as P_{t_0} and the identical task is performed again to extort the pattern at time t and is signified as P_t.

Steps involved	**Overall Process**
Step 1	Self-assertively create an underlying populace of people by methods for an emblematic portrayal procedure. Gauge the wellness of the people by assessing the bond vitality of the applicant arrangements portrayed by them.
Step 2	Pick two or three people from the present populace by methods for a customary roulette wheel choice administrator.
Step 3	The picked people are recombined to deliver two or three posterity by methods for the in part coordinated crossover as per the crossover rates. Gauge the wellness of the two posterity by assessing the bond vitality of the competitor arrangements portrayed by them.
Step 4	Place on record the level of development, or the level of hardship, on their wellness esteems because of the crossover function.
Step 5	Process the regular emblematic mutation of the two posterity and gauge their wellness by ascertaining the bond vitality of the competitor arrangements described by them.
Step 6	Place on record the level of advance, or the level of hardship, of their wellness esteems by virtue of the mutation function.
Step 7	Allocate the resultant people into another populace pool. On the off chance that the populace measure isn't achieved, return to stage 2.
Step 8	Optimize the crossover and mutation rates according to the particular standards.
Step 9	Check the halting basis. Finish up the genetic pursuit methodology and select the best competitor arrangement over time as the last arrangement if the ceasing model is met. Or the consequences will be severe, advance to the succeeding age with the old populace, and returned to step 2.

Fig. 2 Overall step by step processes taken place in the Hybrid Genetic Algorithm (HGA)

7 Patterns Matching Using SCI

The similarity between the two patterns is detected with the help of the similarity computation index (SCI) which is expressed by means of the following Eq. 1.

$$SCI_{xy} = \varphi \times SCI_{xy}^{odd} + (1 - \varphi) \times SCI_{xy}^{even} \quad (1)$$

where, SCI_{xy}^{odd} represents the item, and SCI_{xy}^{even} signifies the linguistic term similarity between two patterns $p_{t_o}^{x}$ and p_{t}^{y} and ϕ characterizes the weighting argument from 0 to 1. The item SCI_{xy}^{odd} and the linguistic term SCI_{xy}^{even} similarity are evaluated by means of the following Eqs. 2 and 3 respectively.

$$SCI_{xy}^{odd} = \sum_{pt=1}^{l} \frac{\left(W_{odd} \times S_{xypt}^{odd}\right)}{W_{odd}} \tag{2}$$

$$SCI_{xy}^{even} = \sum_{pt=1}^{l^L} \frac{\left(W_{even} \times S_{xypt}^{even}\right)}{W_{even}} \tag{3}$$

In Eqs. (2) and (3) the values of SCI_{xy}^{odd} and SCI_{xy}^{even} are estimated in accordance with the position of the pattern which are described as either odd or even. In Eq. (2), l represents the maximal item length of patterns $p_{t_o}^x$ and p_t^y, W_{odd} signifies the weight in the odd position. In Eq. (3), l^L depicts the maximal linguistic term length of patterns $p_{t_o}^x$ and p_t^y, W_{even} corresponds to the weight in the even position. The item and linguistic terms similarity which occur in the pattern's odd and even position are characterized as, S_{xypt}^{odd} and S_{xypt}^{even} respectively. The item similarity and the linguistic term similarity algorithm are efficiently used to evaluate both the item similarity and linguistic similarity and the calculation is concisely presented in [28].

8 Significant Pattern Evaluation

Subsequent to the change patterns association, the significant patterns are estimated in accordance with the user specific minimum support threshold value β. Further particularly, the patterns similarity value is evaluated by their support and $MinSCI_y^t$ values. The emerging patterns significance is estimated by means of the following Eq. 4.

$$P_{xy}^{emerge} = \frac{\left(s(p_t^y) - su(p_{t_o}^x)\right)}{su(p_{t_o}^x)} \tag{4}$$

where the $s(p_t^y)$ and $s(p_{t_o}^x)$ are represented as the supports of pattern p_t^y in D^t and $p_{t_o}^x$ in D^{t_o}. The unexpected change significance is evaluated as per the following Eq. 5.

$$P_{xy}^{unexpect} = SCI_{xy} \times \frac{s(p_t^y)}{s(p_{t_o}^x)} \tag{5}$$

The perished and added patterns significance is estimated by obtaining the product of values between $MinSCI_y^t$ and $MiSCI_x^{t_o}$ value with their support values. The perished and added patterns significance patterns are determined by means of Eqs. 6 and 7 respectively as detailed below.

$$P_{xy}^{perish} = MinSCI_y^t \times su(p_t^y) \tag{6}$$

$$P_{xy}^{add} = MinSCI_x^{t_o} \times s(p_{t_o}^x) \quad (7)$$

In accordance with the above procedure, the significant change patterns are extracted from the databases which facilitate the easy comprehension of the variation in the conduct of the patients.

9 Experimental Results and Discussion

The contemporary technique is performed in the working platform of the MATLAB R2011a (Version 7.12). The epoch-making technique is executed by conducting a series of tests employing diverse dataset values from the UCI dataset. The novel technique elegantly discharges the function of extracting the patterns at two diverse time intervals by means of the Hybrid Genetic Algorithm based Fuzzy Time Interval sequential pattern mining (HGA FTI). Results are displayed based on the criteria of threshold value β which is set by the user (one length frequent patterns) with the modern FTIs. The residual patterns are extracted by means of an optimization approach known as the Hybrid Genetic Algorithm thanks to the fact they consist of further unrefined data. In the long run, the patterns are harmonized with the Similarity Computation Index (SCI) values based on which the patterns are categorized into three different patterns such as the e-Emerging patterns, e-unexpected patterns and the e-Added/Perished patterns.

In the document, an earnest effort is made to categorize the divergence in the conduct of the patient. Thus, three medical datasets such as the Iris, Lung cancer and Pima Indian Diabetes (PID) dataset values are achieved for the varying minimum support threshold values by means of the optimization approaches such as the Hybrid Genetic Algorithm and the Genetic Algorithm were shown in "Table 1".

"Table 2" efficiently illustrates the change Patterns achieved for the minimum support threshold value of 0.1 for the Iris data taken at two different time intervals t_0

Table 1 Datasets obtained at varying threshold values through HGA and GA

HGA				GA			
Minimum support value	Data1 (Iris)	Data2 (lung cancer)	Data3 (PID)	Minimum support value	Data1 (Iris)	Data2 (lung cancer)	Data3 (PID)
0.1	0.8	0.85	0.9	0.1	0.7	0.734	0.78
0.2	0.75	0.85	0.9	0.2	0.71	0.715	0.76
0.3	0.74	0.84	0.87	0.3	0.64	0.674	0.77
0.4	0.8	0.82	0.86	0.4	0.68	0.652	0.76
0.5	0.6	0.81	0.8644	0.5	0.6	0.649	0.7644

Table 2 Change patterns obtained for minimum support threshold value of 0.1 for Iris data taken at two time intervals t_0 and t

Change patterns	At time t_0			At time t		
	E-Emerging patterns	E-Unexpected change	E-Added/Perished patterns	E-Emerging patterns	E-Unexpected change	E-Added/Perished patterns
Original	10	10	10	6	6	6
Predicted	12	10	8	7	5	6

Table 3 Change patterns obtained for minimum support threshold value of 0.1 for lung cancer data taken at two time intervals t_0 and t

Change patterns	At time t_0			At time t		
	E-Emerging patterns	E-Unexpected change	E-Perished patterns	E-Emerging patterns	E-Unexpected change	E-Perished patterns
Original	3	5	4	2	3	2
Predicted	4	3	5	3	3	1

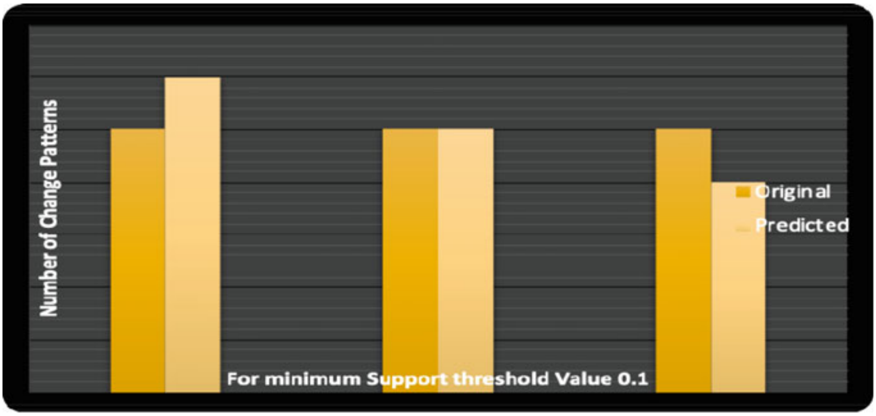


Fig. 3 Change patterns obtained at minimum support threshold value of 0.1 for Iris at time t_0

and t. The values tabulated in "Table 3" are beautifully demonstrated in Figs. 3 and 4 which exhibit the superior outcomes for the predicted values.

"Table 3" efficiently tabulates the change patterns achieved for the minimum support threshold value of 0.1 for the lung cancer data taken at two diverse time intervals t_0 and t with the corresponding values colourfully pictured Figs. 5 and 6 which demonstrate the superior upshots for the predicted values.

"Table 4" elegantly exhibits the change patterns achieved for the minimum support threshold value of 0.1 for PID (Pima Indian Diabetes) data taken at two different time intervals t_0 and t with the values effectively demonstrated in Figs. 7 and 8 showing superior outcomes for the predicted values.

Figures 3, 4, 5, 6, 7 and 8 effectively illustrate the sequential change Patterns achieved at the minimum support threshold value of 0.1 for Iris, Lung Cancer and PID (Pima Indian Diabetes) taken at two diverse time intervals t_0 and t. Results are categorized as Emerging Patterns, Unexpected change for both the time interval t_0 and t. It is crystal clear that the E-emerging patterns achieved for all the predicted datasets are greater in numbers vis-a-vis those of the original datasets, thereby emphasizing the superlative efficiency of the innovative approach.

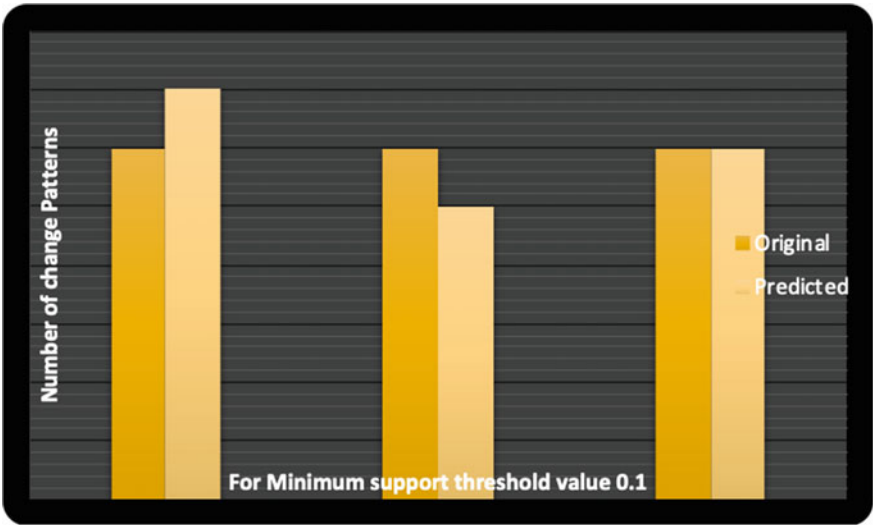


Fig. 4 Change patterns obtained at minimum support threshold value of 0.1 for Iris at time t

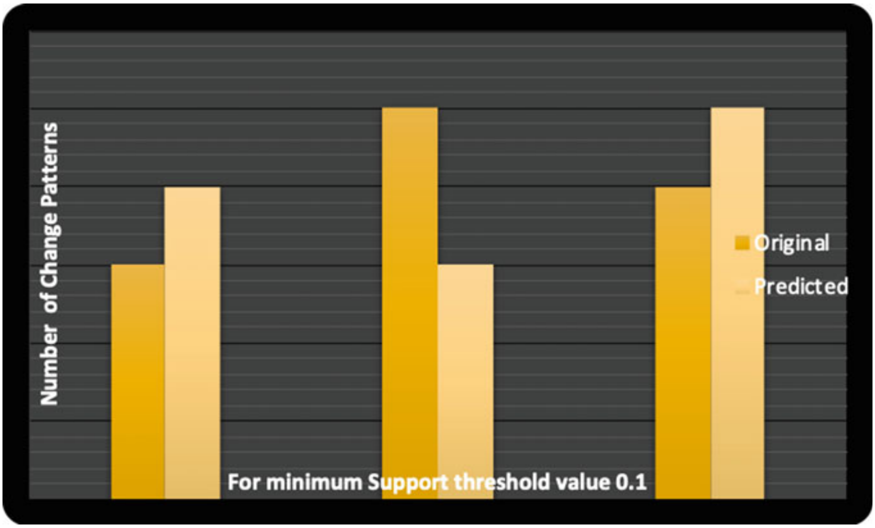


Fig. 5 Change patterns obtained threshold value of 0.1 for lung cancer at time t_0

10 Conclusion

The mounting enthusiasm in the doctor-patient relationship invests its attention on the precise segments of the corresponding association which have a telling impact on the health condition of the patients. With the intention of attaining the medical

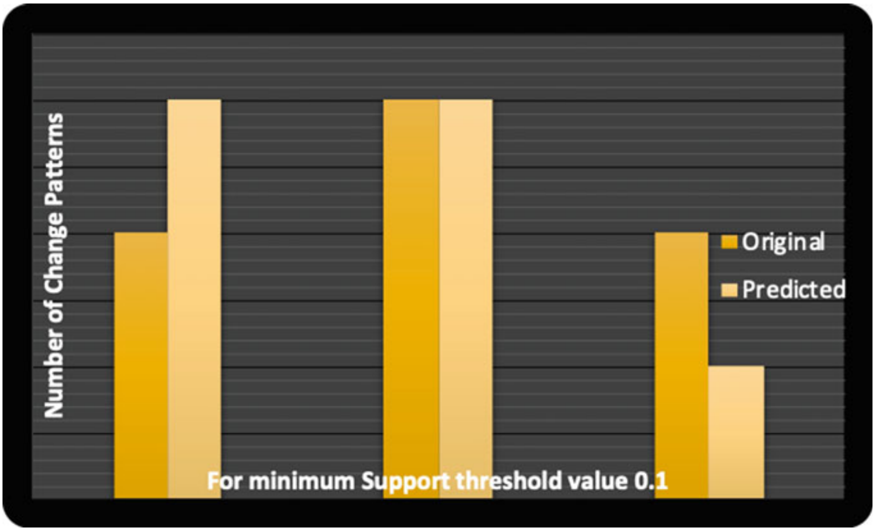


Fig. 6 Change patterns obtained at minimum support threshold value of 0.1 for lung cancer at time t

Table 4 Change patterns obtained for threshold value of 0.1 for PID (Pima Indian Diabetes) data taken at two time intervals t_0 and t

Change patterns	At time t_0		At time t	
	Emerging patterns	Unexpected change	Emerging patterns	Unexpected change
Original	25	25	12	15
Predicted	21	29	14	13

behavioural variations of the patients, two change patterns at diverse time intervals are gathered and processed with the help of the Hybrid genetic algorithm (HGA) based fuzzy time interval sequential pattern mining method. For the purpose, the new-fangled approach performs the following four functions such as the fuzzy time interval sequential pattern mining employing the HGA algorithm, the Patterns matching employing the SCI, the Patterns classification based on SCI value and the Significant pattern evaluation. The extracted patterns are classified into four diverse categories such as the perished patterns, added patterns, unexpected changes, and the emerging patterns and are harmonized by means of the Similarity Computation Index (SCI) values. In the long run, the significant patterns are attained which are used to categorize the divergence in the behaviour pattern of the patients concerned. The milestone method is examined with the help of MATLAB and the efficiency in performance is evaluated.

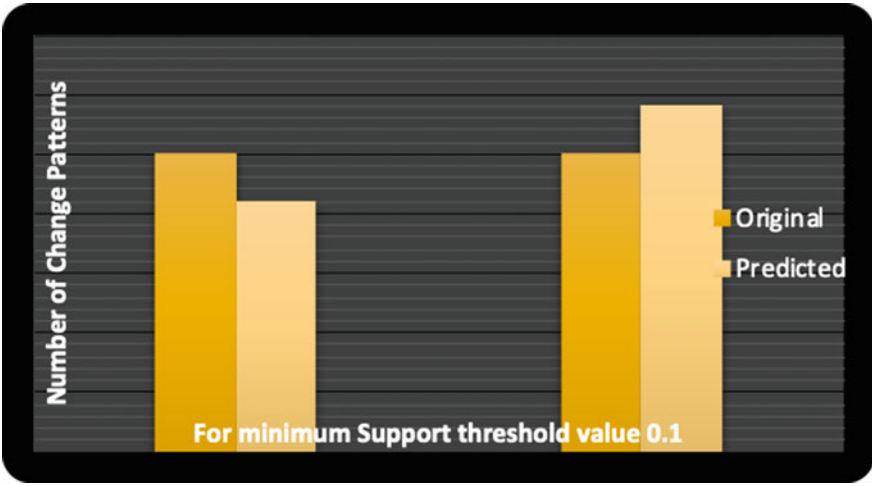


Fig. 7 Change patterns obtained at minimum support threshold value of 0.1 for PID (Pima Indian Diabetes) data at time t_0

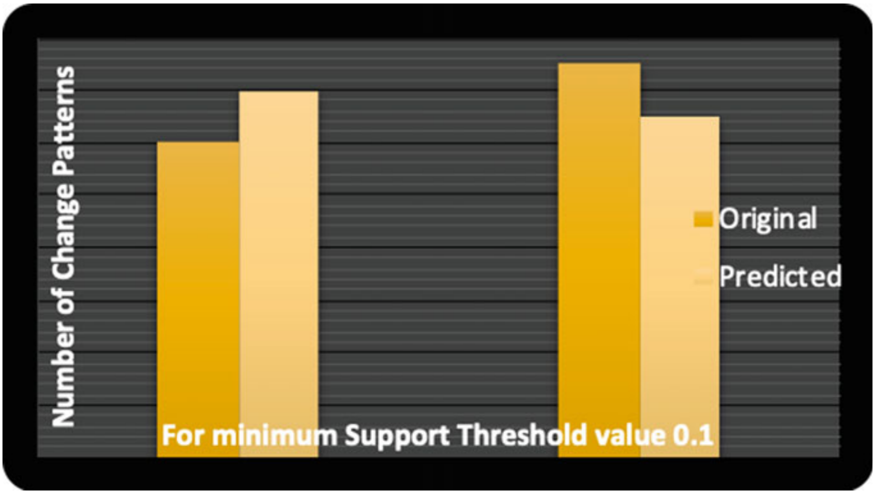


Fig. 8 Change Patterns obtained at minimum support threshold value of 0.1 for PID (Pima Indian Diabetes) at time t

References

1. Bhatnagar, A., Jadye, S.P., Nagar, M.M.: Data mining techniques and distinct applications: a literature review. Int. J. Eng. Res. Technol. (IJERT) **1**(9), 1–2 (2012)
2. Ouda, M.A., Salem, S.A., Ali, I.A., Saad, E.-S.M.: Privacy-preserving data mining (PPDM) method for horizontally partitioned data. Int. J. Comput. Sci. Iss. **9**(1), 339 (2012)

3. Zabihi, F., Ramezan, M., Pedram, M.M., Memariani, A.: Rule extraction for blood donators with fuzzy sequential pattern mining. J. Math. Comput. Sci. **2**(1), 37–43 (2011)
4. Lin, N.P., Hao, W.-H., Chen, H.-J., Chueh, H.-E., Chang, C.-I.: Discover sequential patterns in incremental database. Int. J. Comput. **1**(4), 196–201 (2007)
5. Li, H.-F., Ho, C.-C., Chen, H.-S., Lee, S.-Y.: A single-scan algorithm for mining sequential patterns from data streams. Int. J. Innov. Comput. Inf. Control **8**(3), 1799–1820 (2012)
6. Guyet, T., Quiniou, R.: Extracting temporal patterns from interval-based sequences. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Spain, pp. 1306–1311 (2011)
7. Mabroukeh, N.R., Ezeife: A taxonomy of sequential pattern mining algorithms. ACM Comput. Surv. **43**(1), 1–41 (2010)
8. Hirate, Yu., Yamana, H.: Generalized sequential pattern mining with item intervals. J. Comput. **1**(3), 51–60 (2006)
9. Lin, M.-Y., Lee, S.-Y.: Improving the efficiency of interactive sequential pattern mining by incremental pattern discovery. In: Proceedings of the International Conference on System Sciences, Hawaii, pp. 1–8 (2003)
10. Soliman, A.F., Ebrahim, G.A., Mohammed, H.K.: Collective sequential pattern mining in distributed evolving data streams. In: Proceedings of the International Conference on Innovation and Information Management, Singapore, vol. 36, no. 2, pp. 141–148 (2012)
11. Yang, S.-Y., Chao, C.-M., Chen, P.-Z., Sun, C.-H.: Incremental mining of closed sequential patterns in multiple data streams. J. Netw. **6**(5), 728–735 (2011)
12. Sunita, B.A., Lobo, L.M.R.J.: Data mining in educational system using WEKA. In: Proceedings of the International Conference on Emerging Technology Trends (ICETT), New York, pp. 20–25 (2011)
13. Zheng, Z., Zhao, Y., Zuo, Z., Cao, L.: Negative-GSP: an efficient method for mining negative sequential patterns. In: Proceedings of the Australasian Data Mining Conference, Melbourne, pp. 63–67 (2009)
14. Vijayalakshmi, S., Mohan, V., Suresh Raja, S.: Mining constraint-based multidimensional frequent sequential pattern in web logs. Eur. J. Sci. Res. **36**(3), 480–490 (2009)
15. Motameni, H., Rokny, H.A., Pedram, M.M.: Using sequential pattern mining in discovery DNA sequences contain gap. Am. J. Sci. Res. **14**(4), 72–78 (2011)
16. Pogula, S., Dandu, S.: PTP-mine: range based mining of transitional patterns in transaction databases. Glob. J. Comput. Sci. Technol. **12**(2), 21–28 (2012)
17. Mary Gladence, L., Ravi, T.: Heart disease prediction and treatment suggestion. Res. J. Pharm. Biol. Chem. Sci. **7**(2), 1274–1279 (2016). ISSN: 0975-8585
18. Kandpal, K.C., Agnihotri, R.: SBLOCK—a closed sequential pattern mining algorithm. Int. J. Comput. Appl. Eng. Sci. **1**(3), 296–299 (2011)
19. Mary Gladence, L., Ravi, T., Karthi, M.: Heart disease prediction using Naïve Bayes classifier-sequential pattern mining. Int. J. Appl. Eng. Res. **9**(21), 8593–8602 (2014). ISSN 0973-4562
20. Chang, J.H.: Mining weighted sequential patterns in a sequence database with a time-interval weight. Knowl.-Based Syst. **24**, 1–9 (2011)
21. Niranjan, U., Subramanyam, Khanaa: An efficient system based on closed sequential patterns for web recommendations. IJCSI Int. J. Comput. Sci. Iss. **7**(4), 26–34 (2010)
22. Mehta, J., Mehta, R.: Prefix projection: a technique for mining sequential pattern included length and aggregate. Int. J. Appl. Eng. Res. **7**(11), 1557–1561 (2012)
23. Mary Gladence, L., Karthi, M., Maria Anu, V.: A statistical comparison of logistic regression and different Bayes classification methods for machine learning. ARPN J. Eng. Appl. Sci. **10**(14), 5947–5953 (2015). ISSN 1819-6608
24. Mary Gladence, L., Ravi, T., Mistica Dhas, Y.: An enhanced method for disease prediction using ordinal classification-APUOC. J. Pure Appl. MicroBiol. **9**. Special Edition Nov 2015. ISSN:0973-7510
25. Mary Gladence, Ravi, T.: Mining the change of customer behavior with the aid of Similarity Computation Index (SCI) and Genetic Algorithm (GA). In: The International Review on Computers and Software(IRECOS) in November 2013 issue, vol. 8, no. 11, pp. 2552–2561 (2013)

26. Joshi, S., Jadon, Jain: Sequential pattern mining using formal language tools. IJCSI Int. J. Comput. Sci. **9**(2), 316–325 (2012)
27. Deepika, M., Mary Gladence, L., Madhu Keerthana, R.: A review on prediction of breast cancer using various data mining techniques. Res. J. Pharm. Biol. Chem. Sci. **7**(1), 808–814 (2016). ISSN: 0975-8585
28. Huang, T.C.-K.: Mining the change of customer behavior in fuzzy time-interval sequential patterns. Appl. Soft Comput. **12**(3), 1068–1086 (2012)

Dr. L. Mary Gladence working as an Associate Professor in the School of Computing, Sathyabama Institute of Science and Technology, Chennai, India. She finished her Ph.D. in the domain of Data Mining in the year 2017 from the Sathyabama Institute of Science and Technology. Her research interests include deep learning, artificial intelligence, data mining, machine learning, bio computing, data analytics, with more than 50 publications in these areas. She is currently working on a project titled "Cattle form management using RFID tags", funded by Unnat Bharat Abhiyan (UBA), MHRD, Govt of India.

S. Shanmuga Priya has an undergraduate degree in Information Technology from Bharathidasan University Trichy in 2004, and completed her Masters in Information Technology from Sathyabama University Chennai in 2006. She has completed her Ph.D. in the year 2019. She has 14 years of teaching experience. Her areas of interest include Mobile computing, Wireless Networks and Cloud Computing. She has a number of research papers in the field of Cloud Computing. Security and image processing.

Ms. E. Brumancia joined Sathyabama Institute of Science and Technology in 2008, as an Assistant Professor in Department of Information Technology. She received her M.E. Computer Science and Engineering from Sathyabama University. Currently she is pursuing research in Data Fusion. Her area of interest includes Data Mining, Wireless Sensor networks and Internet of Things. She has published more than 26 research papers in International National Journals and Conferences. 1 Book chapter in Springer. She is a reviewer for Information Sciences and Cluster Computing. Also received an award for "High Impact Factor" Journal publication.

Soft Computing Techniques for Medical Diagnosis, Prognosis and Treatment

Surabhi Adhikari, Surendrabikram Thapa, and Awishkar Ghimire

Abstract With the rapid advancement in the fields of computation and deep learning, the use cases of artificial intelligence in healthcare are blooming more than any time in history. In past years, it was supposed that only doctors and medical practitioners should handle the decisions in healthcare systems. With the rise of machine learning, the tables have turned and dependency on algorithms to make support systems for healthcare has increased. Various AI predictive models have been built for the prediction of diseases at an early stage. Not only this, but data science is also used in a lot of areas of healthcare ranging from summarization of clinical data to intelligent predictive models. However, the work in developing a decision-support system for healthcare is still in the infancy state. Most of the conventional decision-support systems are based on hard computing which requires exactly state analytic models and does not have any place for approximation and uncertainty. Soft computing, being an approach that imitates the human mind to reason and learns in an environment of uncertainty and impression, helps to provide an optimal solution through its nature of adaptivity and knowledge. Various studies have shown that models which extensively used soft computing methodologies, for example, fuzzy logic, ANN, Genetic Algorithms, etc. were able to present clinical observations and inferences in a way that better helped doctors in decision making. There are various applications in the medical field like summarization of clinical text, activity monitoring, development of adaptive disease management systems where soft computing can be used. The book chapter can discuss the prevailing practices, comparative analysis of the necessity of soft computing over more prevalent hard computing techniques, and future directions for the application of soft computing in the healthcare decision-support system.

Keywords Soft-computing · Health care systems · Fuzzy logic · Artificial neural networks · Probabilistic reasoning · Computer-aided diagnosis

---

S. Adhikari · S. Thapa (✉) · A. Ghimire
Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_17

1 Introduction

Artificial intelligence is broadly defined as the programming of machines to enable them to perform tasks like humans. It is an area where machines learn to solve problems mimicking human cognition. Artificial intelligence simulates human intelligence in learning, perceiving, planning, and decision making [1]. Artificial intelligence devices are made to learn the surroundings and act accordingly that also improves their performance with experiences. The main implementations of artificial intelligence and the closely related subfields include computer vision, natural language processing, human-computer interaction, and human-agent interaction. These fields have been leveraged in various sectors such as biomedicine, environment, education, social sciences, finance and economics, cybersecurity, automotive industry, government, law, and most importantly healthcare to improve their domains extensively. Artificial intelligence includes hardware and software both to give human-like decisions on problems related to several fields. With regard to software, artificial intelligence extensively uses algorithms, such as the Artificial Neural Network (ANN) that mimics the human brain cells. The neuron is activated by stimuli, which in AI is the weighted channels and activated functions that work together to form predictions accordingly. These neural networks generate outputs according to the environmental stimulus. Various learning algorithms are present for all sorts of tasks, such as supervised learning, unsupervised learning, reinforcement learning, ensemble learning, and deep learning. Similarly, from a hardware perspective, Artificial intelligence mostly deals with devices that implement the mentioned algorithms on a physical platform to optimize the performance measures and automate problem-solving and decision-making tasks.

As the digital and technological advancements are culminating to reach their zenith, a lot of data is being produced. Data is ubiquitous. Data is produced and consumed at the same time on a very large scale. Data is massively present in the form of text, images, videos, numbers, etc. that can be exploited to reach conclusions, and again the knowledge obtained can be transferred to other tasks. Big data is thus produced and utilized for human development and enhancing human society making human lives easier.

1.1 Healthcare Data

Healthcare data comprises any data that is related with the patient's medical history, drugs given to patients, progression or impedance of the disease, presence of a benign or malignant microorganism in the body, choice of lifestyle, and any other details giving information on the physical or mental state of people. Healthcare data can broadly be divided into two categories, structured and unstructured data. Structured data contains well-ordered data such as the name of the patient along with their medical records including various parameters. For example, in a tumor patient, the size of the tumor, spread rate of the tumor, the amount of drugs taken, can be structured data. Unstructured data, on the other hand, give patients' medical history or records in a not-so-well organized manner. These include clinical notes of doctors, medical prescriptions, audio recordings of the patients, etc. Broadly speaking, healthcare data is enormously present in the form of clinical text, images, and electronic records.

These data can be appropriately utilized for clinical diagnosis, medical history analysis, and prognosis of various diseases. Although utilizing these data and building predictive AI models does not replace real doctors, it is seen that they give significant assistance to the doctors in the entire process. Various learning methods and artificial intelligence techniques can be utilized to deal with the specific type of data. Apart from the regular disease identification and analyzing medical history, AI can also be exploited to bring advances in biomedicine and bioinformation to gain a deeper understanding of different microorganisms, their evolution, and their impacts on the human body. Since, it is a monotonous task for humans to analyze such a large amount of data, with the help of AI, the process becomes a lot easier and saves a significant amount of time. Biomedical literature can be well exploited taking help from AI to get a better understanding of the present scenarios and past developments in biomedicine and healthcare.

1.2 Types of AI in Healthcare

The healthcare data can, thus, be exploited by using supervised learning, unsupervised learning, natural language processing, and deep learning to reach a decision with regard to biomedicine and healthcare. This section gives insights into using different techniques to solve problems in healthcare.

1.2.1 Machine Learning

Machine learning is the most common form of AI used in several fields. Machine learning usually involves supervised and unsupervised learning to make predictions and classifications to data. It usually involves training a large dataset containing several features to come into a conclusion. This can broadly be said as supervised learning. Unsupervised learning, on the other hand, does not take labeled data. It tries to find a distinction between data and finds a correct group for the data. Both supervised and unsupervised learning are utilized in healthcare systems for effective health services to the patients.

The most widely used field in healthcare where machine learning is used is precision medicine. It involves which drug effectively works on a patient given the context and treatment methodologies in a patient. Supervised learning techniques use labeled data to come to conclusions such as the presence of a disease, or if a person will get the disease in the future. It helps clinicians and doctors in the prediction of a disease or other features of a potentially harmful disease by utilizing the past history of the patients.

Unsupervised learning, on the other hand, that majorly involves clustering and principal component analysis (PCA), is used to group patients with similar symptoms or traits. It doesn't conclude into an outcome such as the prediction of a disease or onset of any ailment. But it groups patients and hence the analysis of several patients' medical history becomes easier. Mostly, K-means clustering is used for this purpose. On the other hand, PCA is mostly used for dimensionality reduction. When the features of the patients are recorded in larger dimensions, like the gene history or the number of genes, PCA is used to project the data to only a few components preserving the necessary information about the subject. Usually, in healthcare, PCA is used initially to reduce the data to smaller dimensions, and later clustering is used for grouping them into relevant groups.

Machine learning involves hand-crafted features for analyzing the medical data. Since healthcare is a very vast field and data with delicate handcrafted features are not available all the time. Hence, to deal with this limitation, deep learning is widely used in the healthcare sector.

1.2.2 Deep Learning

Deep learning is a more complex form of machine learning that mostly deals with neural network models. They are being extensively used in health care to deal with data that have many levels of features and outcomes to predict a clinical event. Deep learning can be utilized majorly for analyzing medical images. It can be used extensively for dealing with images that even the human eye may miss out. Deep learning is hence a significant way of providing radiologists, cardiologists, and neurologists in dealing with brain images, heart images, and other images of delicate organs and finding if they contain any lesions. Thus, its use in radiomics is expanding to make a clinically relevant distinction in the medical images of the patients [2]. In the past, computer-aided detection (CAD) had been widely used to exploit the features in medical images. Their combination with deep learning models has proved to give very high accuracy, especially in oncology-oriented image analysis [3]. As a significant amount of medical data is in the form of images, deep learning can essentially be used to gain valuable insights of these images for better diagnosis that eventually leads to better healthcare services to the patients.

1.2.3 Natural Language Processing

Natural Language Processing is the field of AI that deals with enabling computers to understand the natural language of humans and interpret them to gain better insights on text data. Natural language processing, which mainly involves speech synthesis, translation, text data analysis, and information extraction, has been widely used in healthcare systems to enable a better understanding of the natural language used by clinicians as well as patients.

NLP can be used to exploit clinical notes of doctors, and any other unstructured medical text data to be able to make decisions accordingly. NLP can be used in biomedical literature as well to understand the relationships of various biomedical components such as proteins, genes, etc. with the human body. NLP can be used to analyze patients' speech and necessary conclusions can be drawn regarding the progression of any neuropsychological disorder. For example, NLP is extensively studied for Alzheimer's disease (AD) patients and their disorders in the speech which is directly associated with the degeneration of speech. NLP can also be used for conversational AI that converses with patients and hence their speech data can be recorded and analyzed for prediction and diagnosis of clinical events, such as depression or neuropsychological ailments like AD or any other forms of dementia that have an effect in a patient's speech. For example, Alexa, developed by Amazon, is providing assistance to elderly people in their daily activities, and hence has become easier to deal with dementia for them.

Similarly, NLP can also be used in the biomedical literature to extract necessary information. Due to the availability of enormous amounts of information in the literature, it is not feasible for researchers and clinicians to go through such a large amount of data. It is, therefore, necessary to exploit NLP tools to be able to extract only the relevant information from medical journals and help in clinical decision making. Named-entity recognition (NER) has been widely used for this purpose. Similarly, NLP is used in biomedical question answering to help gain fast and accurate answers for questions regarding health and medicine [4].

1.2.4 Physical Robots

Physical robots were initially trained to lift, weld, and do other manual jobs in industries and warehouses. They learn easily and perform accurately in their given surrounding environments. In healthcare, physical robots have done essentially amazing tasks in providing assistance to healthcare professionals in hospitals. Sanitization of medical tools, moving beds, cleaning, can be a few examples of several things that are directly linked with enhancing healthcare services.

Due to blooming advancements in AI, robots have been trained to perform surgery on patients. They assist surgeons and nurses in operating a patient. These surgical robots give significant assistance to surgeons in improving the ability to see, form precision incisions, stitch wounds, etc. [5]. Although the important and sensitive decisions are made by human surgeons, robotic surgery such as head and neck surgery, gynaecologic surgery, etc. have widely been practiced.

1.2.5 Examples of AI in Healthcare

Disease diagnosis and treatment has always been a major goal of AI employed in healthcare. Initially, although AI was developing, it wasn't better than human diagnostics. MYCIN, which was among the first AI technology to have been used in healthcare, was developed at Stanford that was used in diagnosing blood-borne diseases [6]. Since then, a lot of institutions and organizations have been actively using AI tools for prediction, diagnosis, and treatment of diseases.

IBM's Watson is used for cancer detection and necessary treatments. It uses a hybrid of cognitive services through vision, speech and language, and machine learning-based data-analysis programs. Similarly, Google Health has been partnering with a lot of health organizations to come up with solutions regarding various health-related issues. Open-source programs, such as TensorFlow, have been widely used in developing several predictive models to help come up with clinical decisions. PathAI is helping pathologists to help employ machine-learning models for cancer diagnosis and develop methods for individualized treatment by using proper drugs.

2 Intelligent Systems for Healthcare Decisions

Intelligence can be defined as the ability to perceive, reason, analyze, calculate, understand natural language, and compare relationships and analogies between several aspects of the surrounding. The findings and knowledge gained can then be used for solving complex problems, comprehending ideas, and making decisions. Intelligence is a combination of reasoning, learning, problem-solving, perception, and linguistic intelligence. Human intelligence is considered the most supreme form of intelligence due to its remarkable potential for complex decision making, analyzing, interpersonal skills, and outstanding aptitude. Humans are thus distinguished as marvellous beings. Due to the amazing progress, humans have made because of their intelligence, such intelligence is being searched for tremendously. Machines are programmed to mimic this human intelligence to enable them to perform tasks like humans. Although it might take several years to be able to replace human intelligence or even come in par with it, technological advancements in intelligence have become notable in this era. Intelligent systems are being developed massively to enhance human intelligence and assist them in various tasks. Intelligent systems are so vast that they have been deployed gigantly in several fields such as meteorology, e-commerce, businesses, finance, and healthcare. Intelligent systems tend to assist humans in complex decision-making tasks to optimize the use of resources available. Intelligent systems are complex and use a wide range of combinations of technologies—artificial intelligence, wireless networking, computer graphics, cybersecurity, natural language processing, embedded and distributed systems, deep learning, etc.

Intelligence systems can broadly be thought of as computer-based approaches for decision making. Intelligent systems are extensively used in transportation to automate the driving process. The way this is done is by learning the surrounding environment as massive data and being able to make decisions accordingly. Intelligent systems, although, is a very vast field with magnificent advances in technology, when utilizing them, they make sure of the proper amount of energy used so they can be utilized sustainably. Intelligent systems use machine learning to learn from the massive amount of data present. Due to ease of storage and communications at high speeds, intelligent systems can be trained to enable them to make decisions accurately as humans do.

Intelligent systems can be used to manage large amounts of prevalent data in the healthcare domain. This enables healthcare professionals to look into the best medical practices and come up with solutions for treating and diagnosing a wide range of rare diseases. Intelligent systems provide sophisticated approaches to visualize the healthcare data and explore AI techniques in decision-making processes in several processes, such as drug development, patients' recovery, and prognosis. Due to human-like accuracy in decision making in clinical fields, Intelligent systems have proved to become an efficient and significant way of providing healthcare services and care delivery to the patients.

Below, we discuss a few intelligent systems currently in practice in healthcare.

2.1 Virtual Assistants in Drug Development

A remarkable advancement in the clinical field is the ability of AI-informed virtual assistants to automate and speed up the drug discovery process. In the drug development process, clinical trials more specifically randomized control trials (RCT) and control trials are conducted and the effects of a specific drug are recorded with time. This medical report is then analyzed to reach a specific conclusion regarding the relationship of the drug with a specific disease. The clinical trials produce a massive amount of data, and analyzing them manually is a very costly process. Through virtual assistants, these data can be accurately explored to come up with drug development ideas within a lesser amount of time than through manual data handling.

These virtual assistants further help healthcare professionals in finding key answers related to their research. Since they are trained on a specific task, virtual assistants excavate through only the necessary data to come up with answers a researcher wants to know. These can further lessen the time a researcher uses to go through several sources and medical journals in order to come up with a specific finding. Hence, virtual assistants are intelligent enough to comprehend the drug effectiveness in a subject, and furthermore, they are able to extract just the right amount of information from the vast medical literature.

2.2 Intelligent Medical Devices

Medical devices provide significant assistance to healthcare professionals as well as patients. Patients, more specifically, are aided in their daily activities to live an easier and better life. There have been developments of several wearable devices, smart recorders, clinical devices, etc. to automate and ease the process of treatment. Examples include intelligent wheelchairs to aid people with disabilities in their mobility. Speech-based wheelchairs that work on instructions given in the form of speech, help people ease their mobility to a much greater extent. Similarly, ADAMM Intelligent Asthma monitoring can be an example of a wearable technology that assists asthma patients by notifying them if they are approaching an asthma attack. It records the symptoms, past data, use of drugs by patients and hence is able to make such decisions. Similarly, for diabetic patients, by use of an AI-powered insulin pump, they are able to have an insight into how much insulin they need in their blood and when they would need it. Sugar spikes in the blood cause rupture of vital organs in the body, hence diabetic patients need to be extra careful, and AI-powered insulin pumps help them in maintaining the records and eventually giving significant predictions. Moreover, several smartwatches have been developed that effectively keep records of the amount of calories burnt, distance walked, etc. to assist people in decision making regarding their health.

2.3 Ambient Healthcare Monitoring System

Ambient healthcare has been essentially important to comprehend the patients' behaviour in several environments such as hospitals, homes, or parks. It is necessary to observe the patients' reactions in different situations. Sensors such as temperature sensors, Carbon monoxide sensor (CO), Carbon dioxide (CO_2) sensor, and oxygen sensors collectively determine a patient's reaction to certain environments [7]. Similarly, monitoring systems for the patients such as their body temperature, heartbeat rate, body weight, blood sugar level, and several other features need to be recorded to aid PA's (Physician's assistants) as a part of their treatment. These sensors, directly connected with mobile devices can help medical professionals in continuous monitoring of the patients' daily activities as well as their ability to adapt with changes. Banerjee and Roy [8] developed a pulse rate detection system. This used a plethysmography process and displayed the output digitally and hence was able to detect the pulse rate in real-time. Gregoski [9] proposed a smartphone-based heart rate monitoring system. Since heart rate and body temperature are significant indicators of the prevalence of any disease in the body, it is essentially important to monitor them to be able to detect any clinical event at any time, and eventually make decisions accordingly.

3 Use Cases of Soft Computing in Basic Sciences and Diagnostics

Soft computing approaches have found to be working greatly with imprecise data [10]. As it is based on approximate models, they are also able to adapt according to problem domains. Their potential to exploit meaningful and significant relationships set in a data set can further be utilized in the diagnosis, prediction, and treatment of many clinical events. Due to imprecise test measurements, uncertainty and randomness on the normal range of test results, incomplete knowledge on biological mechanisms, and missing information in a large number of cases, imperfect data forms a major part of medical data. Because of this reason, it becomes difficult to find out the best mathematical model or direct computational algorithm to manage this imperfect, incomplete, partially true, or approximate data [11]. Soft computing techniques, thus, are becoming extremely popular in the healthcare industry because of their ability to find a good balance by correctly exploring the randomness in bioinformatics and healthcare data. This makes soft computing powerful, reliable, and efficient in a number of medical tasks such as, drug development, understanding the intricate biology, physiology, and life cycles of microorganisms and their effects in human bodies, interaction of biological molecules in diseases, and finding biomarkers.

The use of soft computing techniques in healthcare can be found in areas like basic sciences, clinical disciplines, and diagnostics measures. The use of soft computing techniques and current works in these areas are discussed below.

3.1 Soft Computing in Basic Sciences

Basic sciences can broadly be defined as the study of biology, chemistry, pathology, or bacteriology that are closely related with life and medicine. Biochemistry, for example, requires the study of complex reactions, proteins, nucleotides, and genes' effect on the enzyme activities of each other, which becomes extremely difficult to analyze. Moreover, conventional mathematical models fail to capture the intricacies of these phenomena. For this reason, Neural Networks, Fuzzy Logic, and Genetic Algorithms have been applied in a number of fields [12]. These include pathology, genetics, biochemistry, cytology, biostatistics, histology, etc.

Genomics is another important field in the basic sciences of medicine. Genomic and proteomic data analysis is important for the knowledge of the fundamental factors of human illness problems. Futschik et al. [13] and Catto et al. [14] used a combination of Fuzzy Logic and Neural Network to identify cancer tissue from gene expression data. Statistical analysis did not perform well in this case. Futschik studied fuzzy rules to detect genes that are associated with particular types of cancer. In understanding hereditary diseases, genomics plays a pivotal role. With this regard, Villmann et al. [15] proposed a soft computing technique of fuzzy labeled neural Gas for the classification of patients suffering from a genetic disease, Wilson's disease. A discrete

Fuzzy logic and neural network model along with the Gaussian variant was applied in the study. Microarray gene expression profiling is another field in genomics that studies the clinical diagnosis of diseases. Their cellular states and biological networks are significant in gene expression data. Ho et al. [16] proposed a Fuzzy logic model tuned by Genetic Algorithms to develop an interpretable gene expression classifier by creating a fuzzy rule base for microarray data analysis. This served as a very easy tool for analyzing gene expression profiles.

Similarly, pharmacology, the area of drug development seems to use soft computing in many of its applications. The field involves in detecting any medicinal or therapeutic substance, their composition, along with toxicology and medical applications. Agatonovic-Kustrin et al. [17] proposed a combination of artificial neural network and genetic algorithms for the prediction of corneal permeability of drugs that have structural differences. This model was able to identify the corneal permeability of the given drugs by analyzing their molecular structure. Furthermore, Agatonovic-Kustrin [17] utilized this model for measuring the plasma concentration in breast milk. It is seen that several kinds of drugs are excreted into breast milk in some quantity. In the study, genetic algorithms were used for finding the subset of descriptors that were found best for transferring the drug to breast milk by studying only the molecular structure of the drug.

Cytology, or cellular biology, is another field in biological life sciences that involves cell studies. The cellular structure, interactions with the environment, their life cycle, division, organelles are studied extensively in this field. Ma et al. [18] proposed a cell slice image segmentation. The complex structure of cells and their sliced images, it becomes an extremely strenuous task for segmenting any biological cell slice image. Fuzzy algorithms and artificial neural networks were used for the segmentation of the image morphologically. This is able to detect edges, regional segments, and wavelet transforms that make it very easy for the cytologists in finding the characteristics of a cell and their impact on any environment.

3.2 Soft Computing in Medical Diagnosis

Medical diagnosis is the recognition or identification of an illness or a disease by examining the symptoms and clinical events in a person. Medical diagnosis can closely be connected with anomaly detection. It is one of the important and delicate issues in healthcare. The correct identification of symptoms leads to a correct diagnosis leading to proper treatments. For this reason, computer-aided diagnosis is becoming extremely popular in the healthcare industry. Soft computing techniques are assisting doctors in identifying subtle changes in the body that may be difficult for human doctors to notice. A few use cases in the medical diagnosis using soft computing techniques have been discussed below.

With regard to medical diagnosis, it is seen that soft computing techniques are majorly used in medical images of the human body. Radiology, hence, seems to be the most popular field where soft computing techniques have taken a significant role.

The medical image study mostly involves ultrasound (US), angiography, magnetic resonance imaging (MRI), and computed tomography (CT). MR images are very prone to excessive noise due to equipment or operator performance, or environment. This leads to a major inaccuracy in the segmentation of such images. Shen et al. [19], Meyer-Baese et al. [20], and Wismuller et al. [21] have used neural-fuzzy systems to address MRI related problems. Similarly, Lee et al. [22] have proposed a multi-modal contextual neural network and spatial fuzzy rules for automatically identifying abdominal organs from CT scans slices. This helped solve a major difficulty in this domain due to gray-level similarities of adjacent organs.

Similarly, Shitong et al. [23] proposed an advanced fuzzy cellular neural network (AFCNN) for the identification of liver images with better accuracies. It gave greater accuracy than the cellular neural network by including the fuzzy logic enabling strong endurance to handle the uncertainties in the images. Raja et al. [24] proposed a neuro-fuzzy hybrid system for analyzing ultrasound kidney images. This method is successful in classifying kidney diseases and also helps physicians in forecasting any future aberrance in the kidney in present normal subjects. N-Benamrane et al. [25] proposed a neuro-fuzzy method in identifying tumors in medical images. Along with the neuro-fuzzy model, the proposed system also uses an expert system. The architecture was tested in MRI images of the brain. Similarly, Andre et al. [26] developed a fuzzy system tuned by genetic algorithms to detect breast cancer. The model was used for testing the Wisconsin breast cancer diagnosis and was able to achieve an accuracy of 97%. Verma et al. [27] proposed a combination of neural networks and genetic algorithms to study digital mammograms. It is seen that it is extremely difficult to recognize a breast tumor as malignant or benign from the mammography images. Hence, it is extremely necessary for a computer aided system to help doctors in this purpose. The model proposed by Verma et al. [27] extracts features and the genetic algorithms select the most relevant ones.

4 Soft Computing Techniques in Healthcare Decision Systems

Soft Computing uses approximate models instead of deterministic models to solve real-world computing problems. Many real-world problems are impossible to be defined by exact deterministic mathematical models. This is where soft computing comes to the rescue. Using soft computing we can solve complex problems by modeling the problem using approximation logic and variables. Soft computing is not a single technique, rather it is an umbrella for a group of techniques and ensembles of various techniques. It consists of Artificial Neural Networks, Genetic Algorithms, Evolutionary Computing, Fuzzy Logic, Expert Systems, etc.

The main objective of healthcare can be loosely identified as the diagnosis of a disease and eventually curing the disease in a human patient. Since healthcare problems are tremendously complicated, the correct diagnosis and cure for a particular disease and anything in between most definitely require the attention of a human mind i.e. a doctor. It is near impossible for a deterministic system to solve healthcare problems, primarily because of the large amount of variables and logic that need to be accounted for. Soft computing seems to be the only way out for computational systems to be able to efficiently solve the issues of health care decision systems. Soft computing derives its inspiration from nature, particularly the human mind and it makes utter sense that if soft computing techniques are able to mimic the human mind to a certain extent, we can have computational systems that can proficiently solve healthcare problems and perhaps even completely replace a clinical doctor.

There are various soft computing techniques currently used in the healthcare industry and they are discussed below.

4.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) derive their inspiration from the human mind and closely resembles the learning pattern of the human mind. ANN is considered to be a universal approximator and it learns a function that maps input to output given the training data. ANNs consist of layers of nodes and each layer of the node has weighted connections to the next layer. The first layer of nodes is the input layer and the last layer of nodes is the output layer. The ANNs learn by approximating the weights of the connection from layer to layer using backpropagation and gradient descent (Fig. 1).

The Value at every node is going to be equal to X = activation ($\sum w_{ji,}\ x_j$). Here activation function can be the sigmoid function or the Relu function. Nowadays mostly the Relu function is used as the activation function. The learning happens in the backpropagation step.

ANNs are intensively used in healthcare systems. Biswas et al. [28] has proposed an ANN-based classification algorithm for the diagnosis of swine flu. Their algorithm gives an accuracy of 94% on their test set. Their ANN consists of 10 input nodes i.e. the feature set they have used has a cardinality of 10. The feature set consists of Fast breathing, sore throat, chills, temperature, runny nose, nausea, cough, fatigue, headache, body aches, and the values that they can attain is in the range of (0–4) i.e. none (0) to severe (4). The ANN consists of 1 output layer which says whether the patient with the given feature set has swine flu or not (Yes/No). There is one hidden layer and the number of nodes in the hidden layer is 14. The activation function used by this ANN is the sigmoid function, and it uses backpropagation as the learning algorithm.

Abdalla et al. [29] propose the use of artificial neural networks for the detection of brain tumor given MRI (Magnetic Resonance Imaging) scans of the brains of patients. Their algorithm gives an accuracy of 99% and sensitivity of 98% which is incredibly good. Their Algorithm consists of six steps.

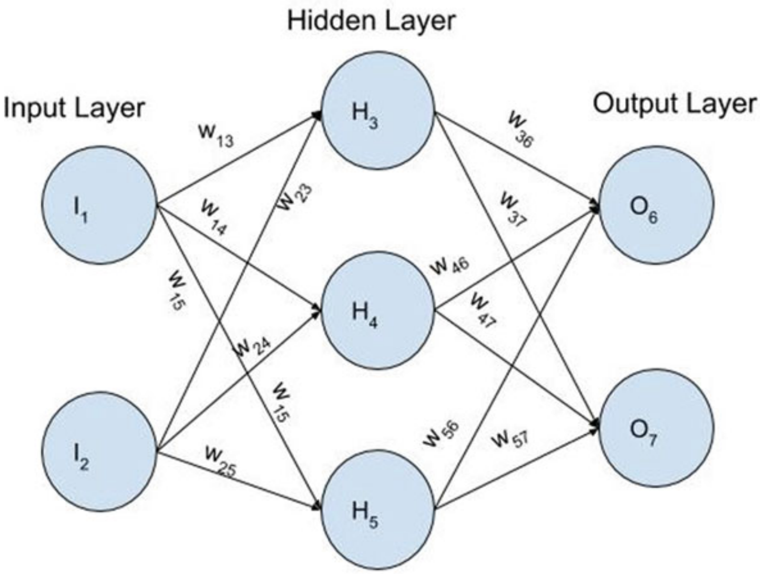


Fig. 1 Artificial neural network

- Database preparation: The data was collected from whole-brain atlas website and the MRI scans were from people above 20 years of age.
- Preprocessing steps: This step consists of reducing the noise in the image and smoothening and sharpening the edges of the MRI image so that there is better feature extraction.
- Image segmentation: The MRI image is then segmented using the threshold method.
- Morphological operations: Then the morphological operations known as erosion and dilation is applied to the MRI image.
- Feature Extraction: This is arguably the most important part of the algorithm. This step used the statistical feature analysis to extract features from the MRI images. The equations of Haralick's features based on the spatial gray level dependence matrix (SGLD) of images computed the features.
- ANN: The features were then put into the ANN. The ANN used the sigmoid function as the activation function.

The above are some examples where ANN has been used to solve healthcare problems. There are many cases in the real world where ANNs have been successfully used to save the life of a human being.

4.2 Fuzzy Logic

A binary logic system is a system in which each variable can have a value of either 0 or 1. A fuzzy logic system can have variables with values ranging from 0 to 1, e.g. 0.67. Using fuzzy logic, we can better approximate real-world problems and solve them accordingly. Fuzzy Logic is extensively used to solve problems in health care. A typical fuzzy logic-based system consists of a fuzzification module, inference system module, knowledge base, and defuzzification module (Fig. 2).

The fuzzification module is responsible for fuzzifying the input i.e. changing the input variables into a range from 0 to 1 accordingly, the inference system and knowledge base consists of a set of rules that work on the fuzzy variables. The defuzzification module defuzzifies the fuzzy variables into a human understandable output which can be further worked with.

Fatima [30] proposes an algorithm based on fuzzy logic for the diagnosis of skin cancer given images of patches on the skin. It uses various image processing techniques and feature extraction techniques to extract information from the image and the classifier used is based on fuzzy logic. The algorithm consists of five steps.

- **Preprocessing**: In this step, the colored image is changed into a grayscale image.
- **Segmentation Image**: In this step, the skin patch is extracted from the rest of the skin.
- **Feature Extraction by using GLCM**: The features are then extracted using the Gray level Dependence Matrix.
- **Diagnostic by using fuzzy logic**: The fuzzy logic in this algorithm uses 9 inputs, and Mamdani's fuzzy logic inference rules.

Simple fuzzy-based logics don't give as high accuracy as neural networks or neural fuzzy systems.

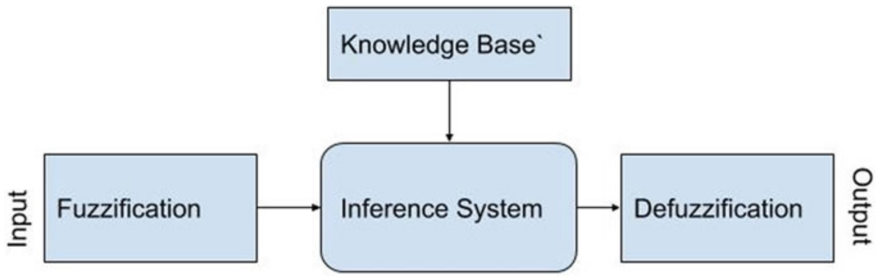


Fig. 2 Working of a fuzzy system

4.3 Genetic Algorithms

Genetic Algorithms are a class of algorithms that gets its inspiration from natural selection and evolution. Genetic Algorithms are mainly used to solve optimization problems. The basic processes of genetic algorithms are as follows:

- **Initialization**: In this step, the initial population is randomly created.
- **Evaluation**: In this step every member of the population is estimated and the performance of the individuals are assessed based on how well they solve the required problem
- **Selection**: In this step, the ones that solve the problem to the greatest extent efficiently is chosen.
- **Crossover**: In this step new offspring are created by incorporating the aspects of the current individuals. After all these processes are over it is expected to create systems that are closer to the requirements that are able to solve the problem. The process is repeated from the second step until a converging point is reached [31].

A breast cancer diagnosis that used neural networks and genetic algorithms was proposed by Khaled et al. [32]. In this model genetic algorithm is used to optimize the weights of the neural network for the highest level of efficiency and accuracy.

5 Further Applications of Soft Computing in Healthcare Decision Making

The usage of soft computing in healthcare is not just limited to medical diagnosis. We can also use Soft Computing techniques for well-being assessment, and health risk assessments like the risk of getting cancer given medical data of a patient, etc. We can also use soft computing for the diagnosis of depression. The diagnosis of depression is a bit different from the diagnosis of other diseases as it is a mental disease and the changes in a brain structure are not so prevalent that depression can be seen in MRI images or other scans of the brain. Yet, we can use soft computing to analyze clinical and questionnaire data of the patient to diagnose depression, as soft computing models are reliable and can be used to solve a lot of real-world problems. Soft computing can also be used to determine the chances of a disease being cured if a patient has a particular disease.

5.1 Fuzzy Logic in Remote Healthcare Monitoring

Hamid et al. [33] propose a method for the healthcare monitoring of the elderly using fuzzy logic. They set up a bunch of sensors and attach it to the person to be monitored or the environment of the person to be monitored. The sensors gather data and then the data goes through a fuzzy logic inference system that outputs the activity that the person is doing. The output includes but not limited to sleeping, washing hands, cleaning, watching TV, bathing, walking, running. This system can be used in elderly people and when abnormal activity is occurring in the monitored person the concerned authorities can be notified.

In the Hamid et al. [33] system there are primarily 3 types of sensors. The first type is a set of microphones that monitors the acoustical environments i.e. the sounds coming from the environment. It also monitors the sounds being produced by the person being monitored. The second type of sensors are wearable sensors that monitor the physiological activities of the person e.g. heart rate, breathing rate, temperature, etc. The third type of sensors are infrared sensors that monitor the environment of the person. This includes smoke detectors etc. Then a master system gathers all the data and uses fuzzy logic to get an output. The data gathered from all these sensors can be imprecise and uncertain and we know that fuzzy systems are able to handle such types of data and give high accuracy results. The advantage of using fuzzy logic for such a system also includes simplicity of design and its ability to deal with complex data from tall these sensors. The data from all the sensors are first analyzed and a correct description is made and the description is numeralized. E.g. If from the first sensor a medium pitched constant sound is coming then the description becomes a vacuum sound and the numerical value for it can be 5. There are different independent algorithms that each of the sensors uses to come up with the description given the raw data. All the inputs are then processed by the fuzzy logic system which uses Mamdani's rules. E.g. If (Sensor 1 has a snoring sound) and (Activity is None) and (heart rate is low) and (Pulse Rate is Low) and (breathing rate is low) Then the Person is (Sleeping).

5.2 Risk Assessment of Cervical Cancer in Women-Based on Convolutional Neural Network

Zahras et al. [34] have used a convolutional neural network to perform the cervical cancer risk assessment. There is a device known as Hinselman's device that magnifies the doctor's view of the cervix using an intense light source and a set of lenses. This acts as the image in the model. The feature set also consists of 32 factors like age, number of sexual partners, number of pregnancies, etc. And there are 4 target variables mentioned in their paper i.e. Biopsy, Schiller, Hinselmann, Cytology. For the Hinselmann target variable, the accuracy is 95.99%. For the Schiller target variable, the accuracy is 95.71%. For the cytology target variable, the accuracy is 97.41% and for the biopsy target variable, the accuracy is 92.69%. Here we can see that researchers have successfully used soft computing techniques risk assessment of cancer.

5.3 Diagnosis of Depression Using Neuro-fuzzy Model of Soft Computing

Subhagata et al. [35] proposed a soft computing model that uses a neural fuzzy hybrid system for the diagnosis of depression in a patient. The feature set used by Subhagata et al. has a cardinality of 14 i.e. there are 14 features that are being considered by this algorithm. Some of the features are loss of appetite, insomnia, indecisiveness, lack of thinking. The output consists of a binary value i.e. Yes/No. Yes if the algorithm diagnoses the input feature set as depressive and No if the algorithm doesn't diagnose the input feature set as depressive (Fig. 3).

The methodology proposed by Subhagata et al. consists of five steps. They are:

- **Data accumulation**: The data is gathered from various hospitals.
- **Principal Component Analysis**: This step uses PCA to extract highly important features so that the data dimension is reduced and then the model can run faster. This is similar to how a real doctor identifies key symptoms first.
- Creating an input vector matrix that comprises significant symptoms as features
- **Fuzzy neural Hybrid Model**: In this part, there is a hybrid model of the fuzzy logic system and a neural network that works in coherence to provide the desired output. The fuzzy system uses Mamdani's laws. The neural network uses backpropagation for training.
- Using real-world depression cases to test the system.

Their algorithm on testing it on real-world depression cases gave an accuracy of 95.50%.

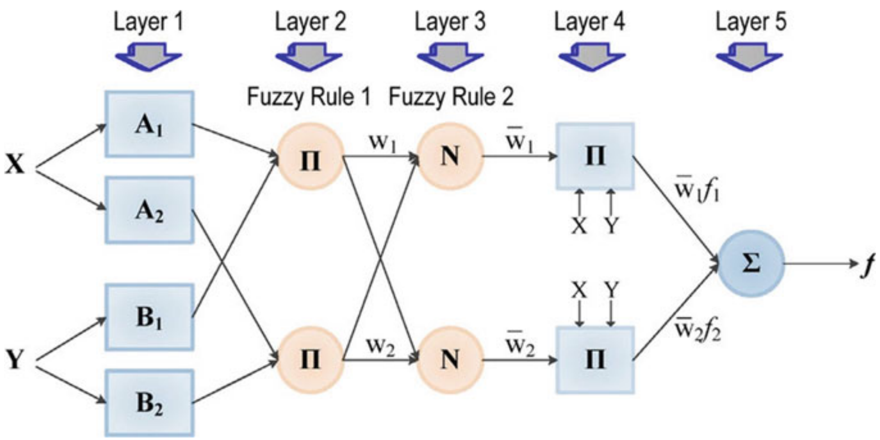


Fig. 3 Schematic diagram of a simple neuro-fuzzy system

6 Hybrid Techniques Used in Healthcare

Soft Computing (SC) is a collection of different techniques and methodologies which aim to harness the tolerance for uncertainty, partial truth, and imprecision to attain tractability, robustness, and low total cost [36]. It is different from hard computing in the sense that hard computing is deterministic and exact whereas soft computing is approximate and deals with problems intuitively and subjectively much like the human mind. Many algorithms in health care use hybrid techniques to solve problems computationally. Hybrid techniques are ubiquitous and every problem that uses soft computing probably has some pre-processing that is done using hard computing. Soft computing and hard computing are not opposite of each other rather they are complementary to each other and the fusion of these two computational disciplines provides humankind with the right tools to tackle the problems of the 21st century. There already exists a large number of systems in which soft computing is together used with hard computing [36].

6.1 Hybrid Solution for Skin Cancer Detection

Aswin et al. [37] proposed a hybrid technique for the detection of skin cancer given images of patches in the skin. The traditional method of diagnosis of cancer cells in the skin is the clinical biopsy, where suspicious parts of the skin are scraped off and sent to the lab for the detection of melanoma [37]. This procedure is time-consuming and expensive. A very cheap and fast method using a hybrid technique is proposed by Aswin et al. Their algorithms preprocess the dermoscopy images using hard computing techniques and the classification is done using soft computing techniques. The classifier they have used is the GA-ANN which is the fusion between genetic algorithm and Artificial Neural Network. In this type of neural network, the weights of the network are optimized to have the optimal value using genetic algorithms. This ensures that the classifier has a high accuracy.

A brief explanation of the steps of the above algorithm is given below.

- Image pre-processing: This is a hard-computing step and uses traditional image processing techniques. In this step, the dermoscopy image is filtered and blurred and smoothened.
- Segmentation: This is also a hard-computing step and the algorithm uses Otsu color threshold segmentation that separates the lesion from the background.
- Feature extraction: In this step features are extracted from the image using statistical techniques and we know that statistical techniques belong to the realm of hard computing.
- Classification using hybrid GA-ANN: The classification is done by using a hybrid fusion of the Genetic Algorithm and Artificial Neural network. In this algorithm, the weights of the neural networks are fine-tuned using genetic algorithms (Fig. 4).

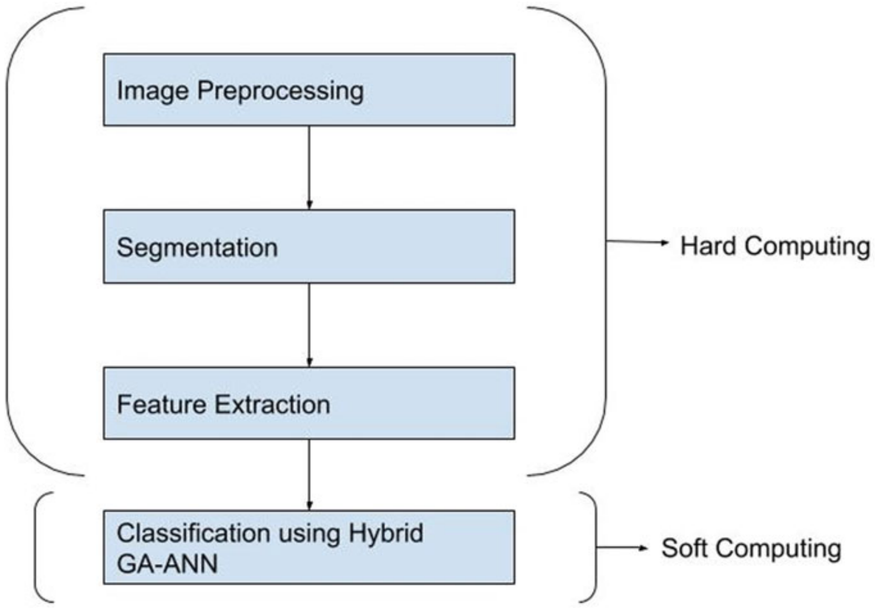


Fig. 4 A schema of a hybrid technique

7 Soft Computing in Clinical Applications

It is seen that soft computing methods are most widely used in clinical sciences. 48% of the soft computing methodologies were used in clinical domains [38]. Soft computing frameworks have been especially used in cardiology, neurology, anesthesia, and rehabilitation. A lot of these fields have applied neural fuzzy mechanisms for clinical assistance to the doctors [39]. The adaptive working of soft computing methodologies has been well utilized in clinical sciences such as controlling blood pressure, unconsciousness, analgesia, etc. A broad discussion of the specific fields and their use of soft computing methodologies have been discussed below.

7.1 Soft Computing in Cardiology

Soft computing is used a lot in Cardiology, particularly while working with ECG. In cardiology medical practitioners extensively use ECG. ECG data are used for diagnosis and monitoring of cardiac functions by enabling the doctors to identify people prone to any cardiovascular event or even death due to any cardiac abnormality. ECG or Electrocardiogram is a graph of the electrical activity of the human heart. In this graph, the y-axis represents the voltage of the electrical activity and the x-axis represents time. Nazmy et al. [40] proposed the classification of ECG signals using

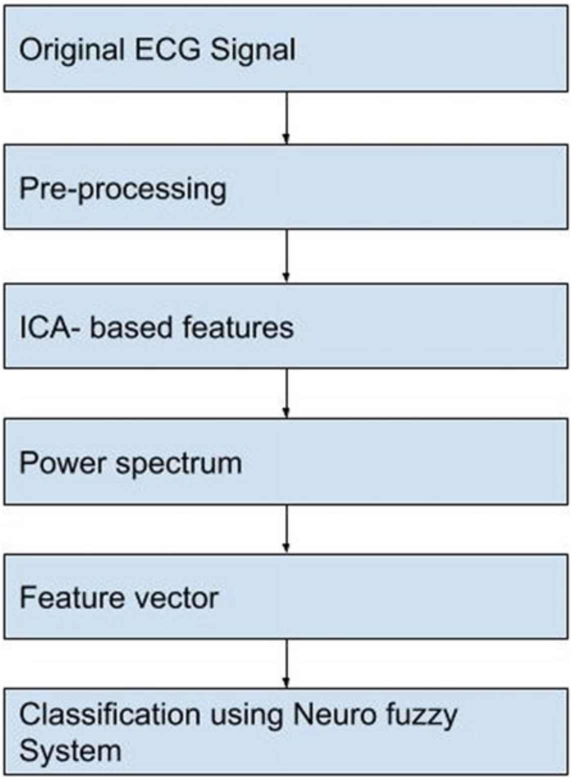


Fig. 5 Classification of ECG signals using adaptive neuro-fuzzy systems

adaptive neuro-fuzzy systems. Their model is shown in Fig. 5 gives an accuracy of more than 97%. Their algorithm consists of 5 steps and a brief description of each of the steps are as follows:

- Preprocessing: Firstly, the data is normalized all the features are standardized to the same level. Since ECG signals may be prone to noise there is the filtering of noise. The signal passes through a high pass filter, a low pass filter, and a Notch filter so that the filtering process is completed and the resulting signal is a noise-free signal.
- ICA based features and Power spectrum: ICA or independent component analysis along with the power spectrum is used for the extraction of useful features which will serve as the input feature vector
- ANFIS: Adaptive Neuro-fuzzy system is used as the classifier in this model. This classifier is used as the diagnostic tool that will help the medical practitioners to diagnose certain heart diseases using ECG signals. This ANFIS uses a hybrid approach of a fuzzy system and neural networks. This ANFIS system has a total of 128 fuzzy rules and one output. The Neural network part of the model uses backpropagation to train. The model can classify six types of ECG signals.

Heart rate signals are considered as a reliable indicator for any cardiac event. Keeping in this mind, Kannathal et al. [41] proposed an adaptive neuro-fuzzy inference system for the detection of heart abnormalities. The model was able to correctly classify cardiac anomalies in 10 different states with an accuracy of 85%. Similarly, Kashihara et al. [42] on an automated drug infusion system. They used a fuzzy-neural model to monitor mean arterial pressure (MAP) in hypotension. Using wavelet transform and fuzzy neural networks, Shyu et al. [43] proposed a method for ventricular premature contraction. Serhatlıoğlu [44] used neuro-fuzzy systems to investigate the consequences of diabetes mellitus on the carotid artery. Acampora et al. [45] recently proposed a combination of fuzzy logic system and fuzzy markup language to study on the ontologies to make the ECG-based knowledge more effective. This was an ECG-based decision-making approach for deducing cardiac health information through heart rate visibility. Using a fuzzy decision support system and genetic algorithms, Paul et al. [46] proposed a system for predicting the risk of heart disease in a currently well person and death risk for those who already had a cardiac abnormality. Similarly, Uyar et al. [47] also proposed a genetic algorithm based recurrent fuzzy neural network decision support system for diagnosing heart diseases.

7.2 Soft Computing in Neurology

Neurology involves the study of the central nervous system and its ontologies. Neurology deals with diagnosis, treatment, prognosis, and investigation of the diseases related to the central, peripheral, and autonomous nervous system [48]. Clinical neurology has immense amounts of imperfect or imprecise data. Hence, soft computing technologies can be applied in this field to aid clinicians for decision making and diagnosis with respect to several neurological disorders. Neurological studies are mostly concerned with electroencephalogram (EEG) analysis, sleep analysis and electromyogram (EMG) analysis. Zhang et al. [49] proposed a nonlinear adaptive fuzzy approximator that enables a non-linear separation of single-sweep evoked potentials. It was also efficient in forecasting non-stationary EEG time-series. Ogulata et al. [50] proposed a neural network method for classifying epilepsy using EEG signals. In epilepsy, the cortical excitability of a person gets completely ruptured. Hence, an accurate method for a correct diagnosis is needed for its treatment and further prognosis. Walczak et al. [51] were able to use artificial neural networks for classifying epilepsy. Their work was the first in this domain.

Schwaibold et al. [52], in their study, compared the classical signal processing approach, artificial neural network, and neuro-fuzzy systems for analysing sleep stages. Neural networks were found to be efficient in pattern recognition due to their robustness. On the other hand, the neuro-fuzzy system was able to grasp the contextual information very well. In their further studies, artificial intelligence in sleep analysis algorithm (ARTISANA) [53] was proposed that automated the entire process of sleep recognition. Khushaba et al. [54] proposed a cognitive fuzzy system for intelligent diagnosis of neurological disorders. It was efficient in assisting the physicians in making decisions with regard to a neuropsychological event in a person. Similarly, Das et al. [55] studied the application of fuzzy systems in hypertension. Hypertension engenders stroke, heart attack, and chronic kidney diseases. Das et al. have done a comparative study on a fuzzy expert system, fuzzy system, and artificial neural network for hypertension diagnosis. Similarly, Sharma et al. [56] proposed a rule-based expert system for neurological disorders, mainly, Alzheimer's, Parkinson's, Migraine, and Meningitis. This proposed system can be used as computer-based assistance for clinical doctors in decision-making processes. Detection of any neurological disorder is extremely essential for a healthy life, as any neurological event can be an early sign of a disorder. Hence, the proposed system is also motivated to provide a home diagnosis system that enables people in identifying a neurological disorder.

7.3 Soft Computing in Medicine and Rehabilitation

Soft computing methodologies can be applied to areas such as critical medicine, physical medicine, and rehabilitation. These areas are mostly associated with therapy, assisting in recuperating a person from a terminal illness, assisting them in living with their difficulties, or getting them habitual with their tasks after a strong medical disorder, such as shock or trauma. This also includes organ transplants and temporary replacement of organ functions by any technical device, such as a pacemaker. In this section, soft computing applications in areas such as intensive care, EEG monitoring, pulmonology, anaesthesia, blood pressure, and respiration regulation, and rehabilitation.

Kwok et al. [57] developed an adaptive neuro-fuzzy inference system (ANFIS) and a multilayer perceptron (MLP) model for ventilator control. Both of these models were able to efficiently model the clinician's decisions. The motivation behind the work was that artificial ventilation of the lungs is of supreme importance to provide oxygen and remove carbon-dioxide from patients whose lungs don't function well. The adaptive fuzzy-neural model was more interpretable than the MLP model. Similarly, Paetz et al. [58] developed a knowledge-based neural network for detecting and helpfully enabling septic shock avoidance for patients in the critical care unit. Belal et al. [59] developed a classifier for categorized pulses into valid and artefacts by implementing a fuzzy inference system. This method was able to monitor pulse oximetry for regulating respiration for neonates and pediatric patients and avoiding false results from probe movement.

Soft computing technologies can also be applied to physical medicine. Physical medicine involves body functional improvement after any injury, congenital disorder, or terminal illness. It is mostly concerned with optimizing the functions of the body and giving a palliative treatment rather than completely removing a disorder or disease. For this purpose, expert training, physical medicines, and physiological modalities are used to ameliorate the cases. Teodorescu et al. [60] developed neuro-fuzzy methods for controlling and diagnosing tremors in a tremor rehabilitation study.

The neuro-fuzzy predictor was a robust and versatile system for coping up with variations of tremor in different individuals. In another study, Deng et al. [61] studied fuzzy-neural models in applying it to a real-time monitoring system for an athlete's daily physical exercise according to the planned schedule. This method successfully improved the efficiency of physical training workload using a computer-aided system. Similarly, in anaesthesia, soft computing can be applied to anaesthesia for controlling and monitoring analgesia, unconsciousness, and blood pressure. Zheng et al. [62] developed a fuzzy logic model for controlling the depth of anaesthesia. This model was successfully able to assess the consciousness level of a patient in a surgery.

7.4 Soft Computing in Other Clinical Areas

Apart from the discussed clinical domains. Soft computing technologies have been extensively applied to other clinical areas such as endocrinology, dermatology, pediatrics, and oncology. In dermatology, the study of skin and skin-related problems, Ubeyli et al. [63] have done notable work in ameliorating the problems related with the detection of erythemato-squamous diseases. They proposed an ANFIS model for differentiating the six shapes of the disease that share the same clinical features. They developed a six ANFIS classifier model for this purpose. Soft computing techniques have been efficiently applied to endocrinology, the specialty associated with internal medicine and hormonal secretions with their interrelationship to physiology and pathology. Endocrinology is also closely related with metabolic functions of the body. The data in this field is hugely incomplete due to very less knowledge, hence soft computing plays a vital role in modeling metabolic systems. Bellazi et al. [64] developed a hybrid neuro-fuzzy method for dynamically modeling metabolic processes. The application also broadly works on intracellular thiamine kinetics.

Similarly, Chen et al. [65] developed a neuro-fuzzy technology for predicting the parathyroid hormone level. In hemodialysis patients, the monitoring of plasma parathyroid hormone is crucially important as their abnormal levels cause renal bone disease. Chen et al. proposed a coactive neuro-fuzzy inference system (CANFIS) for plasma PTH concentration by including clinical parameters. Tung et al. developed a neural fuzzy decision support framework for identifying cancer subtype. Gene expression data was used in their study. Their study was also efficient in detecting pediatric acute lymphoblastic leukemia. In a similar study, Sun et al. [66] used neuro-fuzzy methodologies in oncological applications. The Neuro-fuzzy model was considered a more reliable and accurate method for classifying prostate and breast tumors.

Gastroenterology, dealing with the digestive system and its disorders, is another field of clinical medicine where soft computing methodologies are widely used. Grossi et al. [67] found out in their study that artificial neural networks are well suited for diagnosing gastrointestinal ailments such as chronic pancreatitis, dyspeptic syndrome, or corrosive ulcers. Guler et al. [68] presented a neural network and genetic algorithm method for lung sound classification. For optimizing the parameters of the neural network model, a genetic algorithm was used. Spectral analysis was performed for the chosen breath cycles by a genetic algorithm that was later applied to the neural network. Heckerling et al. [69] proposed a neural network and genetic algorithm for pneumonia predictor variable selection. For the patients with respiratory complaints, the authors used genetic algorithms for searching optimal hidden-layer architectures and eventually diagnosing the problem.

Wu et al. [70] proposed a neural network and genetic algorithm hybrid model for examining patients that had ankle arthrodesis. The model classified the patients with solid arthrodesis with an accuracy of 98.8%. The model showed good performance in analyzing gait patterns in ankle arthrodesis.

8 Conclusion

Soft computing, hence, can bring a very unprecedented shift in the healthcare industry due to its ability to solve complex problems just like a human does. As it emulates the human way of decision making and logical reasoning, it is a perfect tool for solving the drawbacks of the traditional medical decision support systems that are based on traditional AI techniques and statistical or mathematical models. Soft computing is immune to imprecise, uncertain, and incomplete data to manage, manipulate, and mine the data. Healthcare data, being a high variability data and also with a lot of randomness, soft computing works perfectly well. Due to its high adaptability and information processing mechanisms, soft computing is especially useful in handling real-life ambiguous problems. Soft computing methods work better when they are combined rather than just a single method. The hybrid techniques, such as artificial neural network and fuzzy logic, or artificial neural network and genetic algorithms work wonders in manipulating data, extracting features, and providing a meaningful decision about the data. Medical data and records are sensitive in nature. They contain very subtle information that may be important in terms of diagnosis and treatment of a disease. Soft computing techniques help detect these nuances in medical data and assist the clinicians in their decision making. Soft computing techniques can be broadly classified as artificial neural networks, genetic algorithms, and fuzzy logic. All these models are inspired by the way humans have been evolving, developing, and surviving. Genetic algorithms, specifically, imitates the way humans evolve and survive in adverse conditions. The underlying principles lie in the survival of the fittest and natural selection. They work on three genetic operators, selection, crossover, and mutation. Fuzzy logic is mostly used in control tasks. Fuzzy logic involves human language to solve a problem, also known as fuzzy reasoning. It operates on IF-THEN statement schemes. Similarly, artificial neural networks, the imitation of the human neurons, are information carriers. They are interconnected with other neurons, and they collectively work together to pass information signals and come to a decision. Activation functions are the stimuli to these artificial neurons. A combined framework of these soft computing technologies can hence help in decision making just like humans do. This makes possible the use and development of intelligent systems, rule-based expert systems, and computer-aided diagnostics more reliable, robust, and cost and time-efficient. Soft computing techniques are approximations-based models, hence there is no any definite rule in which a system works or a process is modelled. Hence, it might be difficult to understand and choose the right method for solving a problem. Therefore, it is extremely crucial to select the best features of these algorithms, combine them and reach an accurate decision from imperfect, vague, and incomplete data.

References

1. Thapa, S., Adhiikari, S., Ghimire, G., Aditya, A.: Feature selection based twin-support vector machine for the diagnosis of Parkinson's disease. In: 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC). IEEE (2020)
2. Vial, A., Stirling, D., Field, M., Ros, M., Ritz, C., Carolan, M., Holloway, L., Miller, A.A.: The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. Transl. Cancer Res. **7**(3), 803–816 (2018)
3. Thapa, S., Singh, P., Jain, D.K., Bharill, N., Gupta, A., Prasad, M.: Data-driven approach based on feature selection technique for early diagnosis of Alzheimer's disease. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
4. Thapa, S., Adhikari, S., Naseem, U., Singh, P., Bharathy, G., Prasad, M.: Detecting Alzheimer's disease by exploiting linguistic information from Nepali transcript. In: 2020 International Conference on Neural Information Processing (ICONIP), pp. 176–184. Springer (2020)
5. Davenport, T.H., Glaser, J.: Just-in-time delivery comes to knowledge management. Harv. Bus. Rev. **80**(7), 107–111 (2002)
6. Bush, J.: How AI is taking the scut work out of health care. Harv. Bus. Rev. 5 (2018)
7. Islam, M.M., Rahaman, A., Islam, M.R.: Development of smart healthcare monitoring system in IoT environment. SN Comput. Sci. **1**(3) (2020)
8. Banerjee, S., Roy, S.: Design of a photo plethysmography based pulse rate detector. Int. J. Rec. Trends Eng. Res. **2**, 302–306 (2016)
9. Gregoski, M.J., Mueller, M., Vertegel, A., Shaporev, A., Jackson, B.B., Frenzel, R.M., Sprehn, S.M., Treiber, F.A.: Development and validation of a smartphone heart rate acquisition application for health promotion and wellness telehealth applications. Int. J. Telemed. Appl. (2012)
10. Prasad, M., Lin, C.-T., Li, D.-L., Hong, C.-T., Ding, W.-P., Chang, J.-Y.: Soft-boosted self-constructing neural fuzzy inference network. IEEE Trans. Syst. Man. Cybern.: Syst. **47**(3), 584–588 (2015)
11. Gupta, D., Borah, P., Prasad, M.: A fuzzy based Lagrangian twin parametric-margin support vector machine (FLTPMSVM). In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–7. IEEE (2017)
12. Borah, P., Gupta, D., Prasad, M.: Improved 2-norm based fuzzy least squares twin support vector machine. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 412–419. IEEE (2018)
13. Futschik, M.E., Reeve, A., Kasabov, N.: Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. Artif. Intell. Med. **28**(2), 165–189 (2003)
14. Catto, J.W., Linkens, D.A., Abbod, M.F., Chen, M., Burton, J.L., Feeley, K.M., Hamdy, F.C.: Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks. Clin. Cancer Res. **9**(11), 4172–4177 (2003)
15. Villmann, T., Hammer, B., Schleif, F., Geweniger, T., Herrmann, W.: Fuzzy classification by fuzzy labeled neural gas. Neural Netw. **19**(6–7), 772–779 (2006)

16. Ho, S.-Y., Hsieh, C.-H., Chen, H.-M., Huang, H.-L.: Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. Biosystems **85**(3), 165–176 (2006)
17. Agatonovic-Kustrin, S., Ling, L., Tham, S., Alany, R.: Molecular descriptors that influence the amount of drugs transfer into human breast milk. J. Pharm. Biomed. Anal. **29**(1–2), 103–119 (2002)
18. Ma, Y., Dai, R., Li, L., Wu, C.: The state and development of cell image segmentation technology. J. Biomed. Eng. **19**(3), 487–492 (2002)
19. Shen, S., Sandham, W., Granat, M., Sterr, A.: MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization. IEEE Trans. Inf Technol. Biomed. **9**(3), 459–467 (2005)
20. Meyer-Baese, A., Wismueller, A., Lange, O.: Comparison of two exploratory data analysis methods for fMRI: unsupervised clustering versus independent component analysis. IEEE Trans. Inf Technol. Biomed. **8**(3), 387–398 (2004)
21. Wismüller, A., Meyer-Bäse, A., Lange, O., Auer, D., Reiser, M.F., Sumners, D.: Model-free functional MRI analysis based on unsupervised clustering. J. Biomed. Inform. **37**(1), 10–18 (2004)
22. Lee, C.-C., Chung, P.-C., Tsai, H.-M.: Identifying multiple abdominal organs from CT image series using a multimodule contextual neural network and spatial fuzzy rules. IEEE Trans. Inf Technol. Biomed. **7**(3), 208–217 (2003)
23. Wang, S., Fu, D., Xu, M., Hu, D.: Advanced fuzzy cellular neural network: Application to CT liver images. Artif. Intell. Med. **39**(1), 65–77 (2007)
24. Raja, K.B., Madheswaran, M., Thyagarajah, K.: A hybrid fuzzy-neural system for computer-aided diagnosis of ultrasound kidney images using prominent features. J. Med. Syst. **32**(1), 65–83 (2008)
25. Benamrane, N., Aribi, A., Kraoula, L.: Fuzzy neural networks and genetic algorithms for medical images interpretation. In: Geometric Modeling and Imaging–New Trends (GMAI'06), pp. 259–264. IEEE (2006)
26. Pena-Reyes, C.A., Sipper, M.: A fuzzy-genetic approach to breast cancer diagnosis. Artif. Intell. Med. **17**(2), 131–155 (1999)
27. Verma, B., Zakos, J.: A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. IEEE Trans. Inf Technol. Biomed. **5**(1), 46–54 (2001)
28. Biswas, S.K., Baruah, B., Purkayastha, B., Chakraborty, M.: An ANN based classification algorithm for swine flu diagnosis. Int. J. Knowl. Based Comput. Syst. **3**(1), 31–42 (2015)
29. Abdalla, H.E.M., Esmail, M.: Brain tumor detection by using artificial neural network. In: 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), pp. 1–6. IEEE (2018)
30. Ghali, F.: Skin cancer diagnosis by using fuzzy logic and GLCM. J. Phys.: Conf. Ser. (IOP Publishing) **1**, 012020 (2019)
31. Ibrahim, D.: An overview of soft computing. Procedia Comput. Sci. **102**, 34–38 (2016)
32. Alalayah, K.M., Almasani, S.A., Qaid, W.A., Ahmed, I.A.: Breast cancer diagnosis based on genetic algorithms and neural networks. Int. J. Comput. Appl. **975**, 8887
33. Medjahed, H., Dorizzi, B., Istrate, D., Baldinger, J.L., Boudy, J., Bougueroua, L., Dhouib, M.A.: A Fuzzy Logic Approach for Remote Healthcare Monitoring by Learning and Recognizing Human Activities of Daily Living. INTECH Open Access Publisher (2012)
34. Zahras, D., Rustam, Z.: Cervical cancer risk classification based on deep convolutional neural network. In: 2018 International Conference on Applied Information Technology and Innovation (ICAITI), pp. 149–153. IEEE (2018)
35. Chattopadhyay, S.: A neuro-fuzzy approach for the diagnosis of depression. Appl. Comput. Inform. **13**(1), 10–18 (2017)
36. Ovaska, S.J., VanLandingham, H.F., Kamiya, A.: Fusion of soft computing and hard computing in industrial applications: an overview. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **32**(2), 72–79 (2002)

37. Aswin, R., Jaleel, J.A., Salim, S.: Hybrid genetic algorithm—artificial neural network classifier for skin cancer detection. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 1304–1309. IEEE (2014)
38. Yardimci, A.: Soft computing in medicine. Appl. Soft Comput. **9**(3), 1029–1043 (2009)
39. Kiani, M., Andreu-Perez, J., Hagras, H., Papageorgiou, E.I., Prasad, M., Lin, C.-T.: Effective brain connectivity for fNIRS with fuzzy cognitive maps in neuroergonomics. IEEE Trans. Cognit. Dev. Syst. (2019)
40. Nazmy, T., El-Messiry, H., Al-Bokhity, B.: Adaptive neuro-fuzzy inference system for classification of ECG signals. In: 2010 The 7th International Conference on Informatics and Systems (INFOS), pp. 1–6. IEEE (2010)
41. Acharya, R., Kumar, A., Bhat, P., Lim, C., Kannathal, N., Krishnan, S.: Classification of cardiac abnormalities using heart rate signals. Med. Biol. Eng. Comput. **42**(3), 288–293 (2004)
42. Kashihara, K.: Automated drug infusion system based on deep convolutional neural networks. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1653–1657. IEEE (2018)
43. Shyu, L.-Y., Wu, Y.-H., Hu, W.: Using wavelet transform and fuzzy neural network for VPC detection from the Holter ECG. IEEE Trans. Biomed. Eng. **51**(7), 1269–1273 (2004)
44. Serhatlioglu, S., Hardalaç, F., Kiriş, A., Özdemir, H., Yilmaz, T., Güler, İ.: A neurofuzzy classification system for the effects of diabetes mellitus on ophthalmic artery. J. Med. Syst. **28**(2), 167–176 (2004)
45. Acampora, G., Loia, V.: Using FML and fuzzy technology in adaptive ambient intelligence environments. Int. J. Comput. Intell. Res. **1**(1), 171–182 (2005)
46. Paul, A.K., Shill, P.C., Rabin, M.R.I., Akhand, M.: Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 145–150. IEEE (2016)
47. Uyar, K., İlhan, A.: Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Comput. Sci. **120**, 588–593 (2017)
48. Ding, W., Lin, C.-T., Prasad, M., Cao, Z., Wang, J.: A layered-coevolution-based attribute-boosted reduction using adaptive quantum-behavior PSO and its consistent segmentation for neonates brain tissue. IEEE Trans. Fuzzy Syst. **26**(3), 1177–1191 (2017)
49. Zhang, J.-H., Böhme, J.F., Zeng, Y.-J.: A nonlinear adaptive fuzzy approximator technique with its application to prediction of non-stationary EEG dynamics and estimation of single-sweep evoked potentials. Technol. Health Care **13**(1), 1–21 (2005)
50. Oğulata, S.N., Şahin, C., Erol, R.: Neural network-based computer-aided diagnosis in classification of primary generalized epilepsy by EEG signals. J. Med. Syst. **33**(2), 107–112 (2009)
51. Walczak, S., Nowack, W.J.: An artificial neural network approach to diagnosing epilepsy using lateralized bursts of theta EEGs. J. Med. Syst. **25**(1), 9–20 (2001)
52. Schwaibold, M., Schochlin, J., Bolz, A.: Automated sleep stage detection with a classical and a neural learning algorithm–methodological aspects. Biomed. Tech. (Berl) **47**(suppl 1), 318–320 (2002)
53. Schwaibold, M., Schöller, B., Penzel, T., Bolz, A.: Artificial intelligence in sleep analysis (ARTISANA)–modelling visual processes in sleep classification. Biomed. Tech. Biomed. Eng. **46**(5), 129–132 (2001)
54. Khushaba, R.N., Wise, C., Kodagoda, S., Louviere, J., Kahn, B.E., Townsend, C.: Consumer neuroscience: assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. Expert Syst. Appl. **40**(9), 3803–3812 (2013)
55. Das, S., Ghosh, P.K., Kar, S.: Hypertension diagnosis: a comparative study using fuzzy expert system and neuro fuzzy system. In: 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–7. IEEE (2013)
56. Sharma, A.K., Gupta, S.: Neurological disorder diagnosis system. Int. J. Appl. Sci. Eng. Technol. (IJRASET) **2**(5), 296–301 (2014)

57. Kwok, H., Linkens, D.A., Mahfouf, M., Mills, G.H.: Rule-base derivation for intensive care ventilator control using ANFIS. Artif. Intell. Med. **29**(3), 185–201 (2003)
58. Paetz, J.: Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions. Artif. Intell. Med. **28**(2), 207–230 (2003)
59. Bayes, T.: LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philos. Trans. R. Soc. Lond. (53):370–418 (1763)
60. Teodorescu, H.-N.L., Chelaru, M., Kandel, A., Tofan, I., Irimia, M.: Fuzzy methods in tremor assessment, prediction, and rehabilitation. Artif. Intell. Med. **21**(1–3), 107–130 (2001)
61. Deng, W., Sun, X., Fan, X.: Fuzzy control of the physical training intensity based on neural-network. J. Biomed. Eng. **20**(4), 700–703 (2003)
62. Zhang, X.-S., Huang, J.W., Roy, R.J.: Depth of anesthesia control with fuzzy logic. In: Fuzzy Logic in Medicine, pp. 83–112. Springer (2002)
63. Übeylı, E.D., Güler, I.: Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems. Comput. Biol. Med. **35**(5), 421–433 (2005)
64. Bellazzi, R., Guglielmann, R., Ironi, L., Patrini, C.: A hybrid input-output approach to model metabolic systems: an application to intracellular thiamine kinetics. J. Biomed. Inform. **34**(4), 221–248 (2001)
65. Chen, C.-A., Li, Y.-C., Lin, Y.-F., Yu, F.-C., Huang, W.-H., Chiu, J.-S.: Neuro-fuzzy technology as a predictor of parathyroid hormone level in hemodialysis patients. Tohoku J. Exp. Med. **211**(1), 81–87 (2007)
66. Sun, X., Qian, W., Song, D.: Ipsilateral-mammogram computer-aided detection of breast cancer. Comput. Med. Imaging Graph. **28**(3), 151–158 (2004)
67. Grossi, E., Mancini, A., Buscema, M.: International experience on the use of artificial neural networks in gastroenterology. Dig. Liver Dis. **39**(3), 278–285 (2007)
68. Güler, İ., Polat, H., Ergün, U.: Combining neural network and genetic algorithm for prediction of lung sounds. J. Med. Syst. **29**(3), 217–231 (2005)
69. Heckerling, P.S., Gerber, B.S., Tape, T.G., Wigton, R.S.: Use of genetic algorithms for neural networks to predict community-acquired pneumonia. Artif. Intell. Med. **30**(1), 71–84 (2004)
70. Wu, W.-L., Su, F.-C., Cheng, Y.-M., Chou, Y.-L.: Potential of the genetic algorithm neural network in the assessment of gait patterns in ankle arthrodesis. Ann. Biomed. Eng. **29**(1), 83–91 (2001)

Surabhi Adhikari is a final year undergraduate student at the Department of Computer Science and Engineering, Delhi Technological University, New Delhi. She has worked with University of Technology Sydney, Australia on various topics relating to healthcare and stock market analysis. She has also worked with Indian Institute of Technology, Patna, India on various projects in deep learning and bioinformatics. Her research interests are Brain-Computer Interface, machine learning, and AI in healthcare.

Surendrabikram Thapa is a graduate software engineer from the Department of Computer Science and Engineering, Delhi Technological University, New Delhi. He was a visiting scholar to the University of Technology Sydney, Australia in 2019 where he worked on building AI predictive model for the early detection of Alzheimer's Disease. He was an Openlab Summer Student at European Organization for Nuclear Research (CERN), Switzerland in summer 2020 where he worked on Generative Adversarial Networks (GANs). He was also associated with CSIR-CEERI, Pilani on 6-month long project on AI at the edge. His research interests are Brain-Computer Interface, computational learning theory, and Causal discovery.

Awishkar Ghimire is a graduate computer engineer from the Department of Computer Science and Engineering, Delhi Technological University, New Delhi. He works with Amazon India as a software development engineer (SDE). He was also an SDE intern at Amazon, India in 2020. He has published many papers in IEEE, Springer and other reputed venues. His research interests are Computer Vision, Machine Learning, and AI in medicine. He hopes to make affordable clinical diagnosis for common diseases to everyone with the help of AI.

Role of Artificial Intelligence in COVID-19 Pandemic

Abhishek Mehta and Trupti Rathod

Abstract Human services conveyance needs the help of the cutting edge innovations like registering (AI), Internet of Things (IoT), immense insights and Machine Learning to battle and appearance ahead of time against the new sicknesses. We tend to objective to check the situation of AI as an unequivocal innovation to investigate, set up us for obstruction, and battle with COVID-19 (Coronavirus) and opportunity pandemics. Gathered the most current realities with respect to AI for COVID-19, at that point broke down a simply like detect its capacity application for this sickness. We have perceived seven pivotal bundles of AI for the COVID-19 pandemic. This innovation plays a urgent situation to see the bunch of occurrences and to foresee anyplace this infection can affect in the predetermination through gathering and perusing every single past actuality. Human services organizations are a partner degree basic would cherish for dynamic advancements to deal with this infection and encourage them in acquiring right proposals in a timeframe to maintain a strategic distance from its unfurl. Artificial intelligence works in an incredibly master approach to imitate like human insight. It ought to in addition assume a significant job in comprehension and proposing the event of a vaccinum for COVID-19. This outcome driven age is utilized for right screening, breaking down, expectation, and quest for present-day victims and certain future victims. The numerous applications are applied to tracks skill of affirmed, recuperated, and demise cases.

Keywords Artificial intelligence (AI) · Applications · COVID 19 · Healthcare · Technology · Computing · Internet of Things (IoT) · Machine Learning · Healthcare

A. Mehta (✉)
Vadodara, India

Faculty of IT and Computer Science, Parul University, Vadodara, Gujarat, India

T. Rathod
Vidyabharti Trust College of MCA, Gujarat Technological University, Bardoli, Gujarat, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_18

1 Introduction

The presence of bleeding edge PC exhibiting propels, and their appointment, in various parts all around have incited the improvement and quality peril examinations of neighborhood and overall economies. Among which, the clinical field has been particularly changed, where a development in prosperity data emphasized by availability of various advances like Big Data, Machine Learning, Internet of Things (IoT) and Artificial Intelligence (AI) and others has basically helped judicious showing limits. The openness of judicious preparing gadgets and their following use in human administrations worldwide has gotten remarkable changes clinical exercises, singular medicine and the investigation of malady transmission, and such are depended upon to continue improving precision in this field, especially in accuracy of discovering [1]. Watson [2] further notes that these progressions are offering opportunities to judicious assessment through the novel evaluation of recorded and current clinical data. Man-made knowledge perceptive devices are starting at now at use in the selection and assessment of the social protection work power, and are all around also strengthened to fuse stresses of inclusivity and meritocracy [2].

Be that as it may, the vast majority of these registering devices are moderately new and as yet developing concerning the clinical field, and there are various related issues that despite everything should be valued and comprehended. Also, wellbeing experts and related partners have not completely grasped these innovations yet. Consequently, dangerous choices are still intensely dependent upon human-based understandings that are tedious and do not have an extensiveness of information assessments. This is valid in spite of numerous advancements being much more slow when contrasted with choices got from the utilization of customary figuring innovations. On this, it is noticed that lawful structures can assist with directing issues like information assortment and sharing prompting greater receptivity by the wellbeing network. Nonetheless, the absence of normalization of conventions likewise implies that the extent of information to be broke down is restricted to explicit regions. In this manner, results got specifically clinical or potentially geological areas may not make a difference in others.

Hereditary calculations (GA) are an uncommon subclass of a more extensive arrangement of EA procedures. In the accompanying they are particularly featured because of their normal appearance. GA were named and presented by John Holland in the mid-1960s. At that point Lawrence Fogel started to take a shot at transformative programming what's more, Ingo Rechenberg and Hans-Paul Schwefel presented the development methodology. In settling troublesome issues where little is known, their spearheaded work animated the improvement of an expansive class of enhancement techniques (Tettamanzi and Tomassini 2001). Hence the hereditary calculations were concentrated by De Jong and Goldberg. Others, for example, Davis, Eshelman, Forrest, Grefenstette, Koza, Riolo, and Schaffer, to give some examples, GA had been most oftentimes applied to the area of streamlining (Back et al. 1997).

In view of the standards of common development, hereditary calculations are powerful and versatile strategies to tackle search and advancement issues. On account of the power of hereditary calculations, a tremendous intrigue had been pulled in among the analysts [1] everywhere throughout the world. Also, by mimicking a few highlights of organic development, hereditary calculations can fathom issues where customary inquiry and improvement strategies are less viable. Along these lines, hereditary calculations have been exhibited to be promising methods which have been applied to a wide scope of utilization regions. The capacity to apply hereditary calculations to realworld issues has improved essentially over the previous decade. The applications will be presented further in later areas. A neural system is an equal appropriated data preparing structure comprising of various nonlinear handling units called neurons. The neuron works as a scientific processor performing explicit numerical procedure on its contributions to create a yield. It can be prepared to perceive designs and to recognize inadequate examples by copying the human-cerebrum procedures of perceiving data, covering commotion actually and recovering data effectively. As far as displaying, momentous advancement has been made over the most recent couple of decades to improve counterfeit neural systems (ANN). Fake neural systems are firmly interconnected frameworks of so called neurons which have basic conduct, yet when associated they can take care of complex issues. Changes might be made further to upgrade its presentation. Neural systems and fluffy frameworks, generally viewed as components of man-made consciousness, have their deficiencies. A portion of these weaknesses might be survived if fluffy rationale activities are fused into neural systems and neural systems are arranged into fluffy frameworks. In reality, a few writers have just joined fluffy rationale with neural system as neural-fluffy frameworks. It might be another class of processing frameworks gave by the coordination of all these advancing controls for the copying of higher-request intellectual force. They have been applied in different items in various fields.

The utilization of AI-driven calculations for early identification of pandemics is in its development and might be a strong course sooner rather than later to help better readiness. It is normal that as the exactness of these innovations keep on propelling, they will have an increasingly articulated job in advancing the detailing of novel wellbeing strategies. This paper studies how information and AI forms helped in the beginning phases of the identification of the COVID-19 pandemic and gives fundamental supporting proof to grandstand that upgraded information sharing conventions will add to future urban wellbeing strategy globally.

The gigantic degree real factors of COVID-19 casualties may be fused and separated with the guide of using preferred system getting over know computations to higher see the case of viral spread, likewise improve expressive pace and precision, increase novel stunning repairing methodologies, and point of fact see the best inclined individuals essentially subject to revamp innate and physiological characteristics. Motivationally, inside a short interval of time when you consider that COVID-19 scene, preferable structure getting over acknowledge frameworks had been utilized in requested kind of COVID-19 genomes, CRISPR-in a general sense based totally COVID-19 acknowledgment measure, perseverance estimate of uncommon COVID-19 casualties, and going over limit sedate up-and-comers towards COVID-19 (4). Modified mindful strategies can fundamentally get from excellent portrayals of the majority essentially reliant on requested COVID-19 lack of protection. The early remark that developed individuals have a better open door than COVID-19 is tried with the guide of using a latest finding that a growing number of teenagers be badgering by phenomenal COVID-19 appearances, showing a crushing need of an all out chance examination in a general sense reliant on revamp innate and physiological characteristics. This hypothesis ought to be likewise broke down with demanding exploratory plans and dependable period clinical recognitions. As such, natural science (e.g., ACE2 explanation level) and clinical real factors (e.g., age, breath test, viral weight, and perseverance) of COVID-19 casualties with principal clinical conditions may be examined with the guide of using structure getting familiar with methods to now not, now best observe any solid limits (e.g., ACE2) for chance gauge, at any rate besides equivalently complete chance sort and desire for a not too bad planning of advancing issue fix and COVID-19 watchman (Fig. 1). ACE2 innate polymorphism, addressed with the guide of using diverse genetic interpretations in human genome, has been exhibited to influence contamination limiting development (1), proposing a possible inherited tendency to COVID-19 ailment. Along these lines, system getting familiar with appraisal of genetic adjustments from asymptomatic, moderate or crazy COVID-19 casualties may be finished to sort and are expecting individuals essentially reliant on their shortcoming or security from limit COVID-19 tainting, with the guide of using which the structure getting familiar with variation

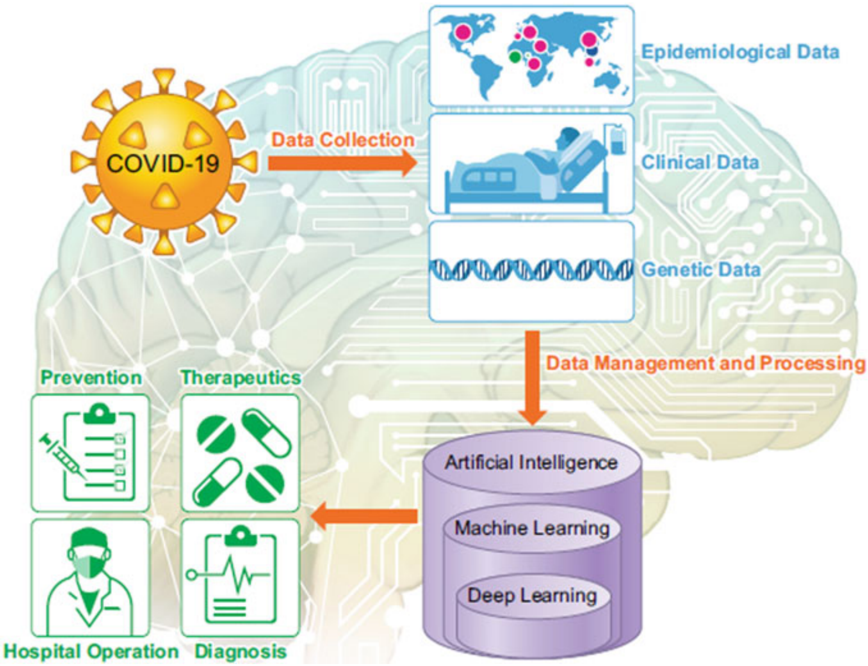


Fig. 1 Application of artificial intelligence and machine learning in the fight against COVID-19

furthermore can restore the ones composed innate structures, including ACE2 polymorphism, of their dynamic path as fundamental capacities with respect to important and foolish assessments (Fig. 1). At the present time, ceaseless undertakings had been made to fabricate novel expressive methods the utilization of system getting familiar with computations. For example, system getting progressively familiar with in a general sense based altogether screening of SARS-CoV-2 test structures the utilization of a CRISPR-mainly based totally disease distinguishing proof contraption got set up with over the top affectability and pace. Neural social order classifiers have been advanced for colossal degree screening of COVID-19 casualties in a general sense reliant on their amazing breath test. Moreover, a significant getting familiar with basically based completely evaluation contraption of thoracic CT photos got worked for modified disclosure and following of COVID-19 casualties after some time (five). Speedy improvement of modified suggestive structures basically reliant on fabricated knowledge and system getting familiar with now not, now best can make a promise to revived characteristic exactness and pace in any case will even watch social protection people with the guide of using diminishing their contacts with COVID-19 casualties (Fig. 1).

2 Premature Detection of the Coronavirus (COVID-19)

The true reprobation gave with the accommodating resource of the usage of the WHO on the scene of the offbeat coronavirus (COVID-19) create as made on the 9 January of 2020, in the wake of having gotten emotions from genuine property in its China Offices. Before long, the contamination advanced beyond time (8 December 2019) at the same time as the primary pack of six patients initially exhorted proper straightforwardly into a Wuhan sanatorium wherein they were managed and discharged. From that treatment date on 31 December 2019, China prosperity real endeavored to choose if or now no longer they were managing a latest kind of disease with out advancement. Along these lines, the postponement in declaring the issue to the WHO China Country Office [3]. Among the bits of knowledge are given to prosperity masters create as that patients who had the symptoms and appearances and signs and signs were totally drawn from the Wuhan city catchment. This extra generally the lockdown of the entire Wuhan city at the 23 January 2020.

Sensitive figuring is a gathering of exciting methods of reasoning, contributed essentially with the guide of utilizing Expert System (ES), Fuzzy Logic (FL), Neural Networks (NN), and Evolutionary Algorithms (EA), which offer versatile realities managing abilities to go to genuine issues. The benefits of using delicate figuring is its ability to experience imprecision, weakness, and fragmentary reality to perform reasonability and vitality on replicating human unique conduct with least exertion (Pal and Ghosh 2004). At the stop of the day, touchy enrolling offers the peril to converse with ambiguity in human reasoning with the weakness, in fact. The most significant touchy figuring procedures are suggested as following.

During the Nineteen Seventies a total hypothesis of proof taking care of realities from several reassets became made with the guide of utilizing Glenn Shafer. It offers a numerical shape to manage and impart to weakness withinside the effect of haziness, imprecision, partial certainty, and absence of realities.

As the pivotal theory of touchy enlisting, fleecy thought process substances numerical weight for the mimicking of the idea and articulation structures. Feathery systems are somewhat valuable in events, for example, significantly convoluted structures notwithstanding in events wherein an extreme affiliation is supported. To control abstract, dubious, sketchy and tangled methods, the fleecy thought process structure might be extremely a dreadful parcel grasped in light of the fact that it introductions a human-like thinking system. Fleecy rationale is an orderly formal multi-regarded thought process idea which utilizes feathery set theory. Its will perhaps formalize the structures of derived reasoning. Cushioned rationale has widely been actualized in unmistakable domains. Soft oversee is one significant model. In feathery oversee, information is depicted with the guide of utilizing semantic components and handle information (IF-Then-guidelines) using those components is purposeful into rule bases. In soft deal with those bases might be applied for canny inferences for controlling purposes. One reason for the achievement of cushioned thought process is that the etymological components, highlights and guidelines enable the draftsman to make an understanding of human information into PC evaluable depictions impeccably. Soft core value bases—if precisely little—might be decoded with the guide of utilizing a pro. This is stand-out to Neural Networks (see ensuing region), that are basically riddle components. The cure of information from a semantic demeanor is an enormous idea in fleecy set hypothesis.

Developmental registering speaks to processing with the utilization of some referred to instruments of advancement as key components in algorithmic plan and execution. A assortment of calculations have been proposed. They all offer a regular applied base of recreating the advancement of singular structures by means of procedures of parent choice, change, hybrid and propagation. The significant one is the hereditary calculations (GAs). GA is one of the stochastic improvement strategies which is mimicking the procedure of characteristic advancement. GA follows similar standards as those in nature (natural selection, Charles Darwin). GA initially was introduced by John Holland as a scholastic research. Be that as it may, today GA end up being one of the most promising methodologies for managing complex frameworks which from the start no one could envision that from a family member humble procedure. GA is relevant to multi-destinations streamlining and can deal with clashes among targets. Subsequently, it is vigorous where various arrangement exist.

In expansion, it is profoundly productive and it is anything but difficult to utilize In recent years, it has been shown that uncertainty may be due to fuzziness rather than chance. Fuzzy logic is considered to be appropriate to deal with the nature of uncertainty in system and human error, which are not included in current reliability theories. The basic theory of fuzzy sets was first introduced by Zadeh (1965). Unlike classical logic which is based on crisp sets of ‘true and false’, fuzzy logic views problems as a degree of ‘truth’, or ‘fuzzy sets of true and false’. Despite the meaning of the word 'fuzzy', fuzzy set theory is not one that permits vagueness. It is a methodology that was developed to obtain an approximate solution where the problems are subject to vague description. In addition, it can help engineers and researchers to tackle uncertainty, and to handle imprecise information in a complex situation. During the past several years, the successful application of fuzzy logic for sol ing complex problems subject to uncertainty has greatly increased and today fuzzy logic plays an important role in various engineering disciplines. In recent years, considerable attention has been devoted to the use of hybrid neural network—fuzzy logic approaches as an alternative for pattern recognition, clustering, and statistical and m thematical modeling. It has been shown that neural network models can be used to construct internal models that capture the presence of fuzzy rules. However, determination of the input structure and number of membership functions for the inputs has been one of the most important issues of fuzzy modeling. Fuzzy logic provides a completely new way of modeling complex and ill-defined systems. The major concept of fuzzy logic is the use of a linguistic variable, that is a variable whose values are words or sentences in a natural or synthetic language. This also leads to the use of fuzzy if-then rules, in which the antecedent and consequents are propositions containing linguistic variables. In recent years, fuzzy logic, or more generally, fuzzy set theory, has been applied extensively in many reservoir characterization studies. This is mainly due to the fact that reservoir geology is mainly a descriptive science which uses mostly uncertain, imprecise, ambiguous and linguistic information. Fuzzy set theory has the ability to deal with Other than the lab ends on the contamination, it's far guided that AI-based counts, outer surface of China, had the alternative and give premature disclosure of the disease yet early than the WHO ascend as instructed, and with inside the end, made sense of how to alert adventurers that is probably at radical risk of being impacted. Inquisitively, the request rises concerning why the style will end up being right now not, now related to the supportive resource of the use of associations in China; head to the wisdom that information sharing principles must be barbed on this heading. Among the a victory associations is BlueDot with the intention of scoured information from records ends, transporter labeling, and animal sickness flare-ups, to predict areas which is in all probability is apparently dependent upon the scene, creating from locale in China [1]. Another business undertaking relationship with some capacities, Metabiota, used the identical system through Big Data assessment to song flight information to capably expect that domain is in all probability in peril to a coronavirus erupt days early than any case ascend as coordinated in any of those nations.

The infection assortment from the ones anticipating device were transferred online to help specialists over the zone in finding an antibody, and to improve their analysis. Following purposeful coordinated efforts among splendid organizations, for the most part in measurements and data sharing, a few degrees of accomplishment in going up against the episode have just been made with the helpful asset of the utilization of a group of specialists drove with the valuable asset of the utilization. This research facility is likewise directed to boast productively positioned the infection in an open record wherein labs acquainted with the helpful asset of the utilization of the WHO can get appropriate of access to strengthen the quest for future immunizations.

Also, the infection might be utilized to grow tests to help discover guardians which are conceivably contaminated and aren't by and by demonstrating any indications and side effects and signs and side effects of the infection, along these lines helping in guaranteeing that the broaden of the sickness is diminished [1].

3 Succinct Review on Transferable Syndrome Outburst in the Year 2020

This is of explicit essentialness because of the reality in the past endemic sicknesses right now are being planned by means of human being introduction to no endemic nations, accordingly cause epidemic episodes. Despite the fact that guidelines right now are being changed to spare you the unfurl of an infirmity to non-endemic nations, a decent arrangement artworks remains needed to find a scourge sooner than its genuine frequency all together that medicinal services strategies might be higher controlled and affected. A PubMed question with the watchwords "("engineered knowledge") OR ("gadget knowledge") OR ("profound knowledge") AND ("infirmity reconnaissance"), likewise helps that present was a developing style in examines in regards to the improvement of AI-basically based absolutely calculations to higher are anticipating the consequences of present day medicinal services data, and thus to are expecting sickness flare-ups ahead of time. In spite of the fact that we do investigate that definite years do now not, at this point evidence any examinations on AI-innovation in infirmity observation, there was incredible examinations finished at the subject of AI-fundamentally based absolutely non-open human services. Figure 2 demonstrates the developing amount of AI-essentially put together thoroughly concentrates with respect to social insurance by means of a look for on PubMed the utilization of the catchphrases "("manufactured insight") AND ("pestilence") spoke to withinside the y-pivot, and the x-hub speaks to year.

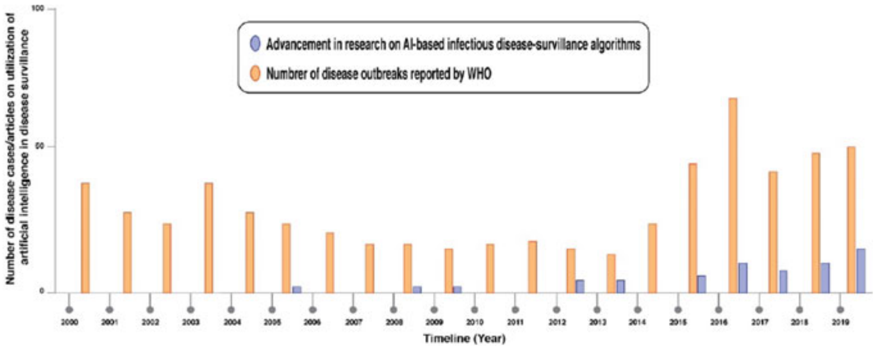


Fig. 2 AI based infectious diseases-surveillance algorithms

4 Applications of Artificial Intelligence in COVID-19 Pandemic

4.1 Premature Detection and Diagnosis of Infection

Man-made intelligence can rapidly investigate unpredictable side effect and other 'warnings' and subsequently caution the patients and the human services specialists [4, 5]. It assists with giving quicker dynamic, which is financially savvy. It assists with building up another conclusion and the board framework for the COVID 19 cases, through helpful calculations.

4.2 Protrusion of Suitcases and Transience

This development can path and guess the possibility of the contamination from the open data, web based life and media stages, about the threats of the sickness and its plausible spread. Supplementary, it envisions the quantity of optimistic cases and death in any territory. Computerized reasoning can lend a hand perceive the most fragile areas, people and countries and acquire checks in like way (Fig. 3).

4.3 Progress of Drugs and Vaccines

PC based knowledge is used for sedating investigate by examining available data on COVID-19. It is significant for a steady transport plan and improvement. This development is used in quickening drug testing continuously, where standard testing takes a ton of time and consequently helps with animating this technique through and through, which may not be possible by a human [4]. It can help with perceiving accommodating meds for the handling of COVID-19 people. It has become an unfathomable resource for suggestive test plans and vaccination headway [4]. Man-made brainpower helps in making inoculations and meds at a ton of snappier rate than anticipated and is furthermore valuable for clinical fundamentals during the improvement of the immune response.

4.4 Tumbling the Work of Healthcare Employees

Because of an abrupt and monstrous increment quantities of people through COVID-19 deadly disease, social insurance experts encompass an extremely elevated outstanding burden. Here, AI is utilized to decrease the outstanding burden of human services laborers [1]. It help in early judgment and giving treatment at a

Fig. 3 General technique for AI applications that assist specialists to recognize the COVID-19 reactions [2]

beginning phase utilizing advanced methodologies and choice knowledge, offer the greatest preparing to understudies and specialists in regards to this new infection [4]. Artificial intelligence can affect expectations unwearied consideration and address increasingly budding difficulties which diminish the outstanding task at hand of the specialists.

5 The Original AI Capability of Bluedot and Metabiota

5.1 Bluedot

For BlueDot to complete expectations on flare-up with respect to irresistible sicknesses, it is predicated intently on AI and device becoming more acquainted with innovation. With these, and the use of different home grown language handling calculations, the business endeavor is prepared to do gather records from a gigantic assortment of reasserts along with data retailers and overall aircraft tagging records [2], beginning its site [2], this beginning up helps that it techniques enormous proceedings sourced from more than 10,000 representative and medium reasserts consistently, beginning more than sixty dialects. It's records place also comprises of realities on masses thickness; questioned from reasserts like nation wide censuses, World Factbook and nation wide records reports. Different reasserts of records that BlueDot is predicated on incorporate the overall Infectious Disease Alert, continuous climate circumstances and bug vectors and creatures' issue supplies. With realities from these kinds of datasets, the business undertaking at that point utilizes sifting apparatus to slim down areas of intrigue [4] and recruit viable grouping rigging to allow the short disclosure of districts that may be showed up as warm spots, bloodless spots and spatial anomalies. This system is also used for the data science and data analytics for the identify the data of coronavirus and also they identify with the python, for the same some NN protocols are used for the same. With the records, the business undertaking at that point applies device becoming acquainted with and home grown language handling innovation to show the machine, and it's far subsequently equipped for transport ordinary markers to its customers, explicitly on occurrences of atypical issue and the possibility stressed and the anticipated areas particularly plausible to appreciate flare-ups. The tutoring of the machine incorporates the utilization of a possibility assessment form that uses the enormous datasets sourced from differing area names to recognize, banner and explain portraying capacity expect the unfurl flare-ups. Crafted by this web-essentially based absolutely fire up had been built up at some phase wherein changed into able to do productively craft forecasts the use of worldwide air visit records to anticipate the overall pathway of the episode.

5.2 Metabiota Metadata

Metabiota metadata depends on innovation calculations to make expectations around irresistible confusion episodes, spreads, intercessions, and event seriousness [5]. Utilizing Neuro-Linguistic Programming (NLP) calculations, this San Francisco basically based absolutely manager assembles monstrous amounts of realities from each true and informal reassets, similar to the ones from organic, financial, political and ecological outskirts, and through unrivaled scientific and perception of frequencies, seriousness, and time of episodes, it might deliver right forecasts [3]. Dissimilar to BlueDot, which scarcely ever depends upon via web-based networking media for realities, Metabioita is situated to here and there, as withinside the current instance of the COVID-19 flare-up, to aggregate and utilize online life realities to make its forecasts. Other than inclusion offices, Metabiota's artworks also points of interest legislative offices, non-income associations, temporary workers, and establishments among various elements who depend upon such records to settle on better-educated decisions in exercises of irresistible issue episodes. It is a result of this that the business delayed its degree to take into account inclusion offices, and these days it has gathered a total issue database. It is currently utilizing present day, prescient innovation like concentrating to make forecasts. subsequent this, while COVID-19 transformed into stated, from beginning to end the utilize of this innovation, it transformed into on the front in foreseeing the accompanying affected wards, aside from Wuhan in which the infection transformed into above all else encountered; this forecast arrived each week ahead of time sooner than the essential examples have been said during the ones goals [3]. This transformed into did through the use of home grown language handling; in which the business transformed into capin a situation to utilize online networking realities from outstanding reassets to tune the unfurl of the infection; henceforth, render considerably more noteworthy right expectations.

6 Conclusion

The situation of Artificial Intelligence (AI) with inside the late identification radical coronavirus is archived on this document from end to end the works as a art of organizations, BlueDot and Metabiota, it shows AI-pushed calculations can deliver additional specific expectations and readings with inside the fate through improved records sharing. The paper helps that improved records sharing activity should be authorized inside the city wellness region simultaneously as standing through the size of privateers and assurance as a result of the sensitive idea of insights happening engineering. happening this, AI techniques illustration beginning well-dressed records reasserts furthermore commencing well-dressed city innovation and that related mechanical ideas, combined with wearable advances, can and should be energized, in light of the fact that it will deliver huge datasets and subsequently extra right forecast and discovery. For this completion, there's a need for the normalization of conventions to rouse verbal trade among devices and all through structures without bargaining records assurance and halting records oversight. The innovative upheaval upon us will see a developing utilization of registering strategies, and as their exactness will increment higher control choices can be delivered; as with inside the instance of pandemics, and could hence bring about its extraordinary situation in city wellness strategy.

References

1. Brickwood, B.: The power of AI in providing rapid at-home assessments of the coronavirus. The Health Europa, 6th March 2020
2. Suthaharan, S.: Supervised learning algorithms. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems, vol. 36, pp. 183–206. Springer, Boston, MA (2016)
3. Tryfona, N., Price, R., Jensen, C.S.: Chapter 3: Conceptual models for spatio-temporal applications. In: Sellis, T.K., et al. (eds.) Spatio-temporal Databases. Lecture Notes in Computer Science, vol. 2520, pp. 79–116. Springer, Berlin, Heidelberg (2003)
4. Lao, K.K., Deb, S., Thampi, S.M., Fong, S.: A novel disease outbreak prediction model for compact spatial-temporal environments. In: Thampi, S., Gelbukh, A., Mukhopadhyay, J. (eds.) Advances in Signal Processing and Intelligent Recognition Systems. Advances in Intelligent Systems and Computing, vol. 264, pp. 439–450. Springer (2014)
5. Zin, T.T., Takahashi, R., Hama, H.: Robust person detection using far infrared camera for image fusion. In: Second International Conference on Innovative Computing, Information and Control, 2007, ICICIC'07, Sept 2007, pp. 310–310 (2007)

Prediction of Transmittable Diseases Rate in a Location Using ARIMA

Varun Totakura and E. Madhusudhana Reddy

Abstract As the effect of COVID-19 is increasing rapidly every day, it becomes very difficult for the survival of many people and there is a high effect on the economic situations of every country which is affected by it. In the same way, it is affecting all states of India and causing an economic crisis. This paper deals with the analysis of the impact caused by COVID-19 on each state in India and also gives an estimated date on which the effect will reduce along with the analysis report of overall India. For the predicting of the effect, we have used Auto-Regressive Integrated Moving Averages (ARIMA) algorithm which has produced Root Mean Squared Error (RMSE) of around 5.89 for some of the states and other with 20.05 due to the data abnormality. The predicted data for each state projected in the figure using the line plots. And the resulted graphs are explained clearly. The accuracy of the proposed model is around 94.6–96.8% for the states with good data and less RMSE values and 80% for the states with abnormal data and high RMSE values. From the produced results of the proposed methodology, the dates of which the effect of COVID-19 will decrease is calculated for the states which have a high number of cases.

Keywords Data analysis · Virus disease outbreak · ARIMA · Forecast · RMSE · Visualization

1 Introduction

The pandemic disease COVID-19 is out spreading rapidly throughout the world. As per the statistics in June, 2020 the number of cases reached around 10.3 million and no. of deaths has reached around 5.3 lakhs. In India, the no. of cases is above 560 thousand and no. of deceased people are above 16,800 and it is still growing. This

V. Totakura (✉)
Tata Consultancy Services, Hyderabad, Telangana, India

E. Madhusudhana Reddy
Department of Computer Science and Engineering, Guru Nanak Institutions Technical Campus, Hyderabad, India

S. Dash et al. (eds.), *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, Studies in Big Data 89, https://doi.org/10.1007/978-3-030-75657-4_19

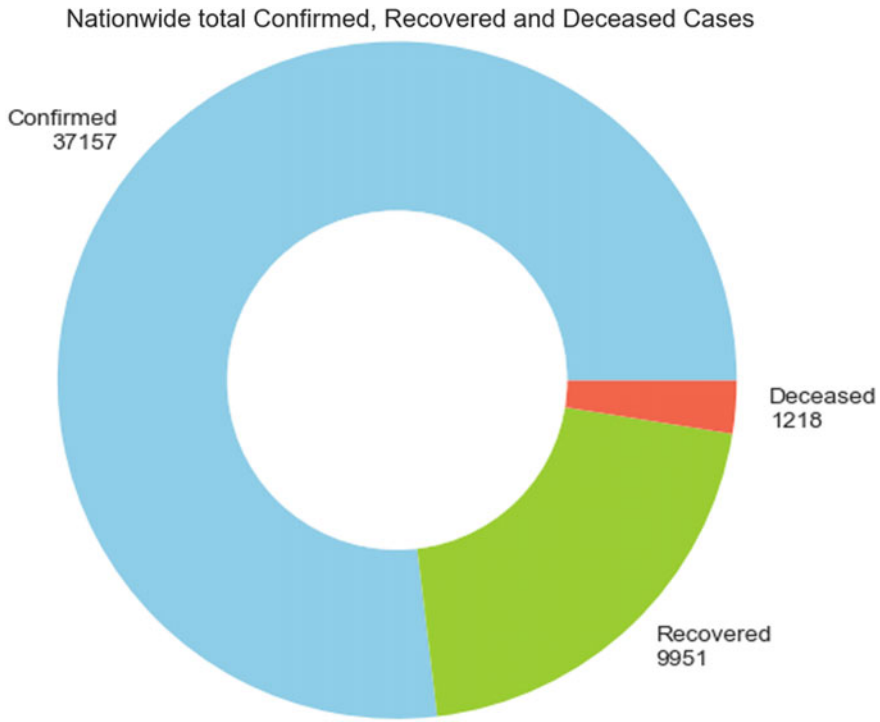


Fig. 1 COVID-19 cases, recovered and deaths in India, as of 1st May 2020

disease has started in 2019 at Wuhan, China and had spread in almost all countries on the globe. The data about the cases in India is shown in a pie chart in Fig. 1. People who are infected with virus will fall sick and experience difficulty in breathing. It spreads from an infected person to other people when the infected person sneezes, coughs or exhales. These droplets are too heavy to hang in the air, and quickly fall on floors or surfaces. The virus can live from 2 to 3 days on plastics, steel and on paper, and on wood it can live up to 5 days. As per the available data it was said that a person who was infected with mild COVID-19, on average can recover in 2 weeks and with severe infection they can recover in 3–6 weeks approximately. Along with the increasing of the cases the recovery rate is also increasing. Throughout the world there are more than 5 million who have recovered from this disease and in India the number crossed 330 thousand. The governments of respective countries and states are following some of the preventive measures that is helping to decrease the rate of spread. In India, the government has imposed a lockdown period for some days to prevent the spread of the virus. During the lockdown period, no person is allowed to commute from one place to another without proper permission and many organizations like IT industries, Schools, Colleges, Universities and many more are closed down. Only few organizations which are related to the essentials of the humans are permitted. Due to the implementation of lockdown there is a significant change

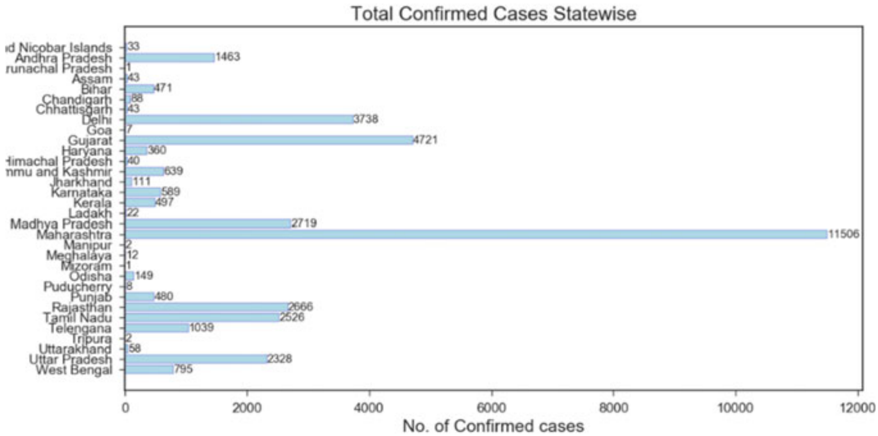


Fig. 2 COVID-19 confirmed cases in the States of India, as of 1st May 2020

in the spreading rate which is falling down. The following bar chart in Fig. 2 shows the number of confirmed cases in each state in India.

At places where people are strictly following the rules and preventive measures imposed by the respective state governments the rate of spread of COVID-19 is falling down or very less, but at places where the rules are broken, the rate is very high. But there is a chance of having an economic crisis and poor people will not be able to buy their essential goods. In keeping all these conditions as a priority there is a need for an analysis to predict the impact of the virus on the people which helps to spread the awareness among people in many ways so that the people start judging the situation in a right way and the government can implement or suspend the rules or preventive measures that are taken against the spread. If the people start to take care by maintaining the social distancing, then the government can permit some organizations to start their work again which help in avoiding an economic crisis. Keeping all the above reasons in mind we have proposed a statistical model to analyze the situation in each state in India.

The proposed statistical methodology will take the time-series data which contains statistical information of no. of cases, recovered, deaths in each state in India as per the records on 15-June-2020 [1]. The time series data will provide an opportunity to forecast future values. Based on the previous time series data and by using the forecasting statistical algorithms the values of the same parameters can be predicted for the future. But the model which was used should be accurate so that there are no wrong decisions made. Forecasting is a tremendous task which is generally performed by higher officials of an organization for the growth of their business and even by the stock market investors to get good amount of profits. The future trend of stocks or the sales will help investors or owners to make the right decisions. The same can be used for the prediction of the impact of COVID-19 in India. We have used ARIMA model in our paper for the forecasting of impact of COVID-19 as it is one of the most used method specifically for time series data. Forecasting a time series can be broadly divided into two types. If the previous values of the time series data are used to predict its future values; then it is called Univariate Time Series Forecasting. And if we consider predictors other than the series to forecast it is called Multi Variate Time Series Forecasting. ARIMA is a forecasting algorithm based on the idea that the information from past values of the time series can alone be used to predict the future values. And to confirm that the predicted data is accurate, various error calculating methods are used.

We have used RMSE method for the calculation of the error rate of our proposed model. Root Mean Square Error (RMSE) can be defined as the standard deviation of the residuals or the prediction errors. Residuals or prediction errors are a measure of how far from the regression line data points are in the normal form (Without noise). It is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Generally, the RMSE is obtained from Mean Squared Error (MSE). The MSE is a commonly used error calculation method for the statistical data. RMSE is just the square root of the MSE. It is probably the most easily interpreted statistic. For the observations that are made by the data available by our model, the RMSE values ranges from 5 to 22 approximately because of data uncertainity. The RMSE value is low for the states for which the data has some kind of trend from the starting but it is high for states in which the number of the cases is abnormal. Further details about the proposed model and the respective calculations are described in below sections.

2 Related Work

Basically the COVID-19 analysis is an important tool for all the governments to make the right decision, the researchers are more attracted to perform analysis in an accurate manner and for the best accurate results. The ARIMA model was used in the prediction of COVID-19 cases in the year 2020, [2] the ARIMA models can be used as an immediate tool for the prediction of time series data for health monitoring systems. It is a good model for short-term forecasting but there should be a good procedure in the process of interpretation. The ARIMA interpreted that there will be around 200,000 cases in Italy. The usage of the ARIMA model is wide ranged, it can be used on the time-series data for the prediction of daily or monthly or even yearly average.

A similar kind of work was performed by Almasarweh and Alwadi [3] on the prediction of the banking stock market data. They have mentioned that the ARIMA model was very useful for the short term analysis with few time series of data. They have used MSE method in calculation of the error value to calculate the accuracy of the model. A graph was plotted for the dataset which tells about the comparison of the banking and index.

The same kind of model was developed for the Nigerian Stock Exchange by Adebiyi et al. [4], Alsharif et al. [5]. The prediction of global solar radiation, by Awajan et al. [6] in the prediction of the stock market data using EMD-HW bagging and Totakura et al. [7] have developed a LSTM Neural Network model for the stock market prediction.

The generalized review about the ARIMA model for the prediction or forecasting of time series data was represented in a paper published by Liu et al. [8]. They have proposed a novel online method using the ARIMA model. They have theoretically proved that their method has produced the most better results from previous fixed ARIMA methods. They have even compared their algorithm with the previous algorithms on Auto Regressive Moving Averages (ARMA) and have mentioned that their model performed well.

A review work on forecasting the electricity price was performed by Weron [9]. The paper will give the information about the electric price forecasting study and also interprets on the directions of the price for a decade.

A generalized forecasting method was given by Taylor and Letham [10]. The first system that they have made to forecast was performed on the data from Facebook. They have used a modular regression model for their first method. And secondly, they have performed a study on tracking the forecasting accuracy of the model. A study was performed using [11] Weka tool for the data visualization and analysis of networking data.

A similar kind of methodology was also proposed by Armstrong [12] and also by Kumar and Anand [13] in prediction of sugarcane production in India. A study was performed on jute production by Das et al. [14].

The prediction of the tourism was performed using the ARIMA model [15]. The paper has interpreted that the tourism in F. Y. R. Macedonia will play a major role in contribution for the country's economy. They have used ARIMA (1, 1, 1) model for the prediction and analysis of international tourism at F. Y. R. Macedonia. The accuracy of the model seems to be good but not perfect as there should be an increase in the accuracy for accurate prediction which will help in making correct decisions for both tourism and economic growth. And due to advantages of ARIMA, it was used by many of the authors [16–21] for predicting the future or forecasting the future values of the statistical data to study and analyze.

3 Methodology

Auto Regressive Integrated Moving Average is a highly used statistical model specifically used for time series analysis. It is a generalized model of Auto Regressive Moving Average model. Basically, these two models are used for understanding and forecasting of the time series data. ARIMA is also mostly applied on non-stationary data. Because, it has an integral part which will help in removing non-stationarity from the data. A stationary time series data is the data in which the vales of the data will not depend on time. The time series which exhibits trends or seasonality are considered to be non-stationary. ARIMA model is the combination of both Auto Regressive (AR) and Moving Average (MA) model. The equation of the ARIMA model can be written as:

$$y't = c + \phi 1y't - 1 + \cdots + \phi py't - p + \theta 1\varepsilon t - 1 + \cdots + \theta q\varepsilon t - q + \varepsilon t$$

The term 'Auto Regressive' in ARIMA gives the interpretation that it is a linear regression model that uses its own lags as predictors. Generally, the Linear regression models will work best when there is no correlation with the predictors and also when they are independent to each other. To make the time series data stationary, the most common approach which is used is the differentiating method. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the time series data, more than one differencing is performed. Then the value of d will become the minimum number which will help to make the time series data stationary. And if the time series is already stationary, then d = 0. 'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors. And 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

$$\underset{\substack{\uparrow \\ \text{AR(p)}}}{(1-\phi 1B-\cdots-\phi pBp)}\ \underset{\substack{\uparrow \\ \text{d differences}}}{(1-B)\,dyt} = c + \underset{\substack{\uparrow \\ \text{MA(q)}}}{(1+\theta 1B+\cdots+\theta qBq)\varepsilon t}$$

After the calculation we get the estimated p, q, and d values which should be given as the parameters to the ARIMA model as shown the Fig. 3 to obtain the Akaike's Information Criterion (AIC) values.

AIC will be helpful in the process of selection of estimators and also to determine the order of the model which will help in getting accurate results. The AIC can be calculated by using the below equation:

$$\text{AIC} = -2\log(L) + 2(p + q + k + 1)$$

The example of the AIC values which are obtained by the ARIMA model when fitted with the COVID-19 "Andhra Pradesh" State data are as shown in Fig. 4. Among the obtained values the least AIC value is taken and the corresponding parameters are given to the parameter list of the ARIMA model.

After obtaining the AIC values the ARIMA model is fitted or trained with the time series data which is the state COVID-19 data. And the results of the model can

```
Examples of parameter combinations for Seasonal ARIMA.
SARIMAX: (0, 0, 1) x (0, 0, 1, 12)
SARIMAX: (0, 0, 1) x (0, 1, 0, 12)
SARIMAX: (0, 1, 0) x (0, 1, 1, 12)
SARIMAX: (0, 1, 0) x (1, 0, 0, 12)
```

Fig. 3 Examples of the parameters p, d, q values given to ARIMA model

```
ARIMA(1, 1, 0)x(0, 0, 0, 12)12 - AIC:382.7930577094712
ARIMA(1, 1, 0)x(0, 0, 1, 12)12 - AIC:1349.8427995020402
ARIMA(1, 1, 0)x(0, 1, 0, 12)12 - AIC:302.45364070581377
ARIMA(1, 1, 0)x(0, 1, 1, 12)12 - AIC:184.47528661517313
ARIMA(1, 1, 0)x(1, 0, 0, 12)12 - AIC:291.3554690920168
ARIMA(1, 1, 0)x(1, 0, 1, 12)12 - AIC:1340.0250958845795
ARIMA(1, 1, 0)x(1, 1, 0, 12)12 - AIC:185.83037137423074
ARIMA(1, 1, 0)x(1, 1, 1, 12)12 - AIC:183.8762486009413
ARIMA(1, 1, 1)x(0, 0, 0, 12)12 - AIC:370.5255667788842
ARIMA(1, 1, 1)x(0, 0, 1, 12)12 - AIC:1217.0176398667315
ARIMA(1, 1, 1)x(0, 1, 0, 12)12 - AIC:291.8271786523849
ARIMA(1, 1, 1)x(0, 1, 1, 12)12 - AIC:174.77358103778263
```

Fig. 4 Examples of the AIC values obtained by ARIMA model

Table 1 Table which shows the result of ARIMA model

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2576	1.165	0.221	0.825	-2.025	2.540
ma.L1	-0.9321	1.169	-0.797	0.425	-3.224	1.359
ma.S.L12	-0.7888	1.660	-0.475	0.635	-4.043	2.466
sigma2	22.6742	29.265	0.775	0.438	-34.685	80.033

be represented as shown in Table 1 which contains the vales of the Auto Regression, Moving Averages and the corresponding coefficient values, standard errors, and many more. These values can be represented in the graphical mode as shown in Fig. 5 which consists of a combination of four graphs which represent the values of the Standard residual, Histogram plus estimated density, Normal Q-Q, and Correlogram graphs. These two figures contain the results that were obtained by training or fitting the ARIMA model with the data from the state Jharkhand, India.

The predictions of the proposed Statistical model that is ARIMA model for each state can be seen in the results discussion section of this paper. Along with these, details of the predicted date from which the impact of COVID-19 may decrease is also displayed using a table. And the error calculation and accuracy of the model is also displayed in the same results section.

4 Comparative Study

The Comparison of prediction by the proposed ARIMA model can be performed on the predictions performed by the ARMA model. The ARMA prediction model was applied to predict the Rain fall by Nugroho and Simanjuntak [22]. They have taken

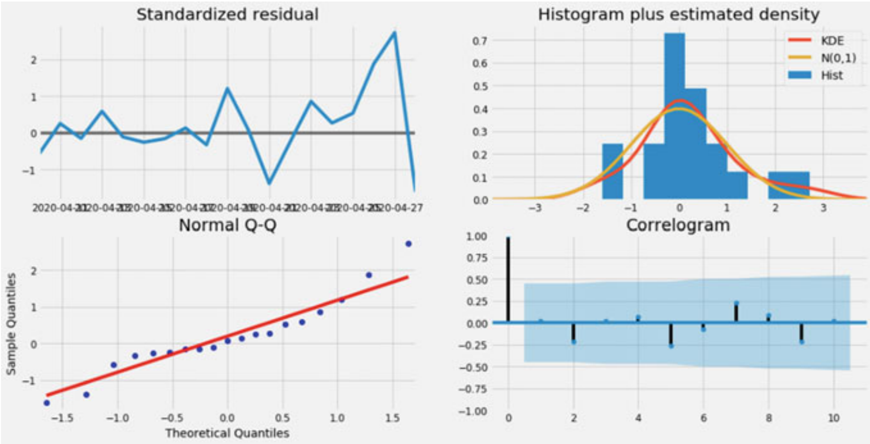


Fig. 5 Graphs which shows the result of ARIMA model

rainfall data of the Republic of Indonesia from 2001 to 2013 month-wise data and fitted to a Univariate Time Series ARMA Model. In their study, their proposed ARMA model had produced results which were not synchronizing with the previous data. The MSE of that model was below 10% for a selected AIC value and had produced good results, but the stationarity of the data need to be manually achieved. As the data which need to be fitted to statistical model especially ARMA, the stationarity of the data plays an important role. If the data is non-stationary, then the results may become extremely variate from the actual result, but when the data is stationary then good accuracy can be achieved and the data can be comparatively matched with the actual data.

A study of rainfall was performed by Sidiq [23] where a method for forecasting of rainfall for a year was developed using ARIMA model on the Indonesia data. The data which was used by the author for training the model ranged from 2011 to 2014. In this study, the author had considered the Mean Absolute Deviation (MAD) method for the calculation of the best fit ARIMA model. After picking the best p, q, d values from the calculation using MAD, the model was fitted with the input data. But, if we check the table mentioned in that paper, there is a huge variation in the predicted data from the previous data. The predicted data was highly normalized which says that the model was under-fitted and has produced normalized data.

The ARIMA model presented in this paper had produced better results when compared with the above mentioned cases. The model was not over fitted or under fitted and the usage of ARIMA will reduce the time consumed to make the input data stationary as the differential part was integrated in the ARIMA model. Table 2, describes as easier way of comparative study for the above mentioned methods.

Table 2 Comparative study with previous methods

Paper	Model	Drawbacks
Nugroho and Simanjuntak [22]	ARMA	Delay due to manual differentiation to make input data stationary
Sidiq [23]	ARIMA	The model was under fitted and had produced highly normalized results

5 Results and Discussion

The proposed ARIMA model has produced good results with the selected p, q and d values. But, as the accuracy of the model depends on the data that is given, for every state data that is given to the model its respective p, q, d values are used according to the least AIC value. The graph obtained using the predicted values of the number of cases and actual values by the proposed statistical model on the time series data of Jammu and Kashmir state was displayed in Fig. 6. It proves that the model was finely fitted to that data as the predicted and the actual data seems to be equal. In the graph displayed below the blue line tells the actual or observed values whereas the red lines give the information of the predicted data.

The prediction of the number of recoveries by the proposed model in the state of Kerala using the data of the number of recoveries from that state is as shown in Fig. 7. In the graph, the blue line is the actual vales of the number of recoveries with the day scale data and the red line is the predicted values that seem to be similar. Thus, we can say that the data of that state has fitted with the model very accurately. But in every data, there will be some noise and actual values may become extreme and the predicted data will follow a trend line.

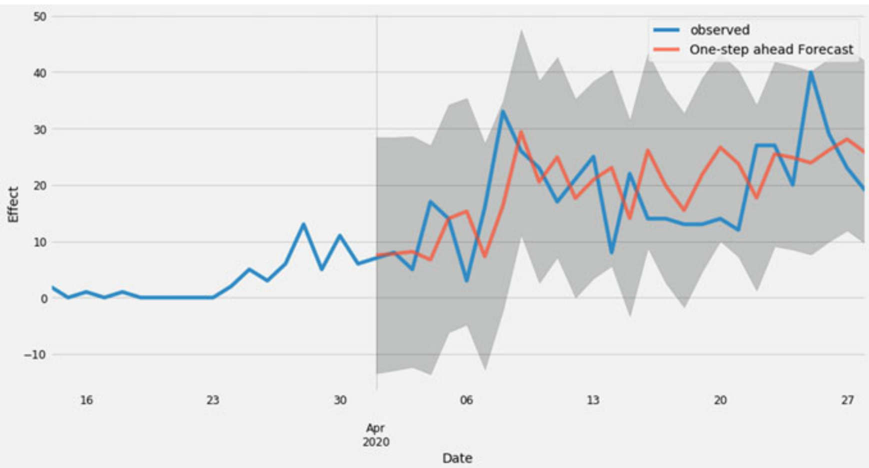


Fig. 6 Prediction of number of cases in Jammu & Kashmir State

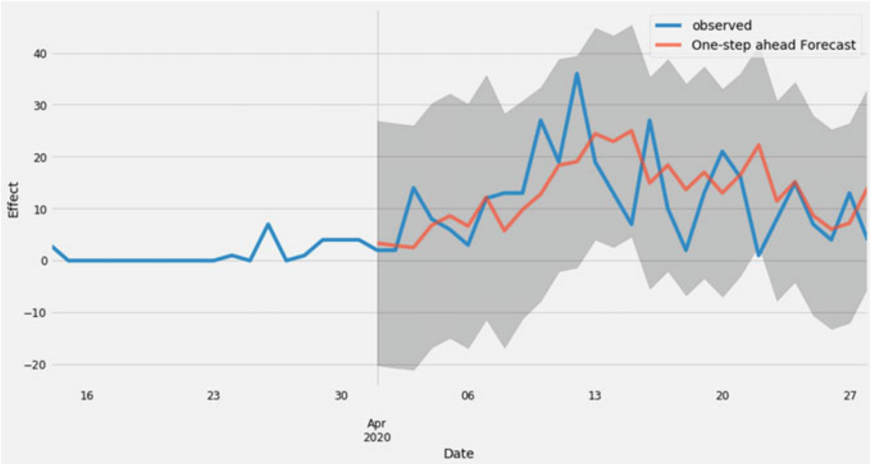


Fig. 7 Prediction of number of recovered in Kerala State

The predictions of the cases and the recovery in the future by the ARIMA model is shown in Fig. 8. Here from the graph we can see that the actual recovery values are displayed with a yellow color line, actual cases are displayed with a blue line, red line gives information about the predicted number of cases in the future and the green line tells about the number of recoveries of Telangana state in the future. From the graph we can interpret that the Telangana seems to have less number of cases in the future and the recoveries will be rising. This seems to be a better change where the COVID-19 spread may reduce gradually. In the following figures, the information

Fig. 8 Prediction of number of cases and recovered in Telangana State

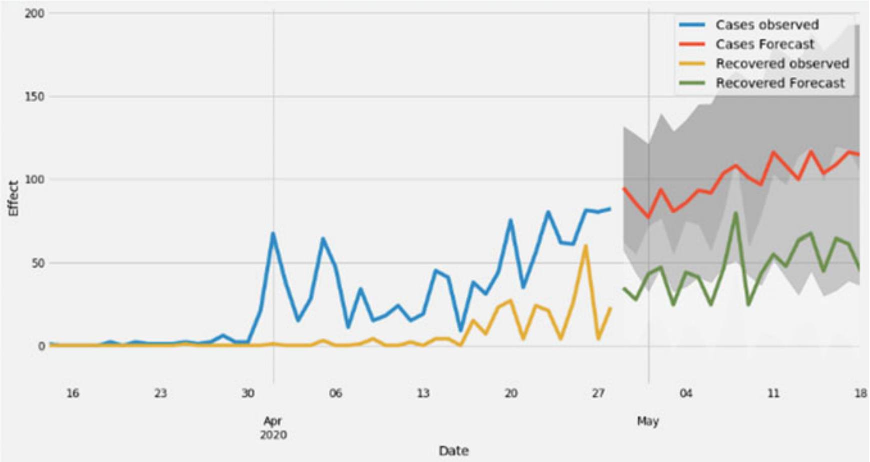


Fig. 9 Prediction of number of cases and recovered in Andhra Pradesh State

which was predicted by the model for other states are displayed. The states and the respective graph is arranged in a format.

In Fig. 9, the predicted number of cases and recovered number of the Andhra Pradesh state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is less than the predicted cases so, there might a possibility that the cases may become very high in coming days.

In Fig. 10, the predicted number of cases and recovered number of the Haryana state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is more than the predicted cases so, there might a possibility that the cases may become very less in coming days.

In Fig. 11, the predicted number of cases and recovered number of the Jammu and Kashmir state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is less than the predicted cases so, there might a possibility that the cases may become very high in coming days.

In Fig. 12, the predicted number of cases and recovered number of the Jharkhand state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number

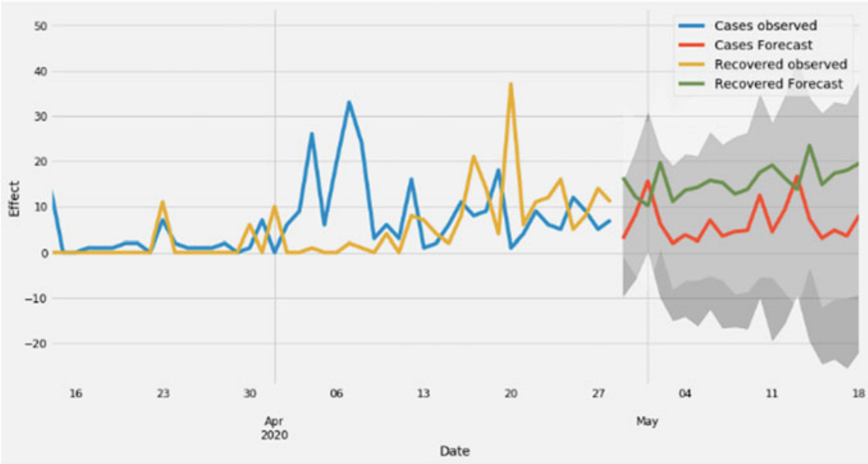


Fig. 10 Prediction of number of cases and recovered in Haryana State

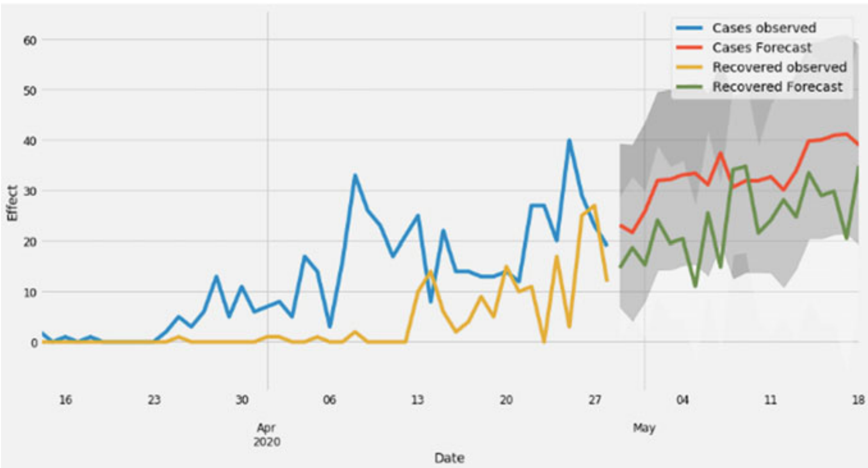


Fig. 11 Prediction of number of cases and recovered in Jammu and Kashmir State

of recoveries. In the figure it is observed that the predicted number of recoveries is less than the predicted cases so, there might a possibility that the cases may become very high in coming days.

In Fig. 13, the predicted number of cases and recovered number of the Karnataka state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is

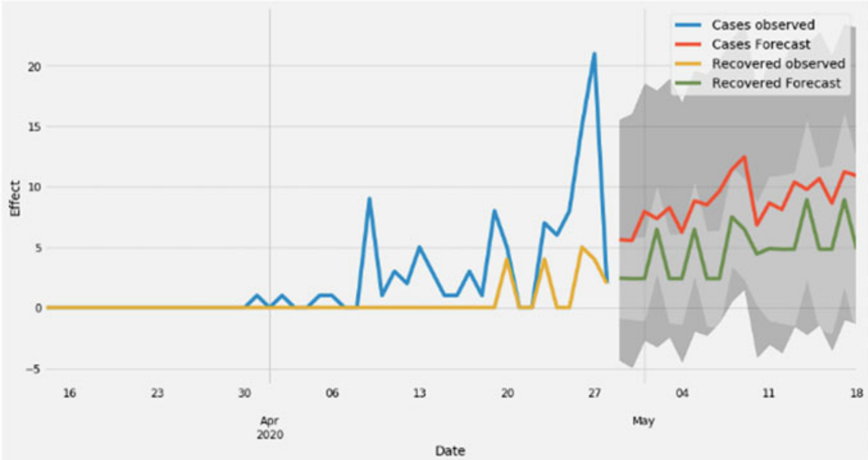


Fig. 12 Prediction of number of cases and recovered in Jharkhand State

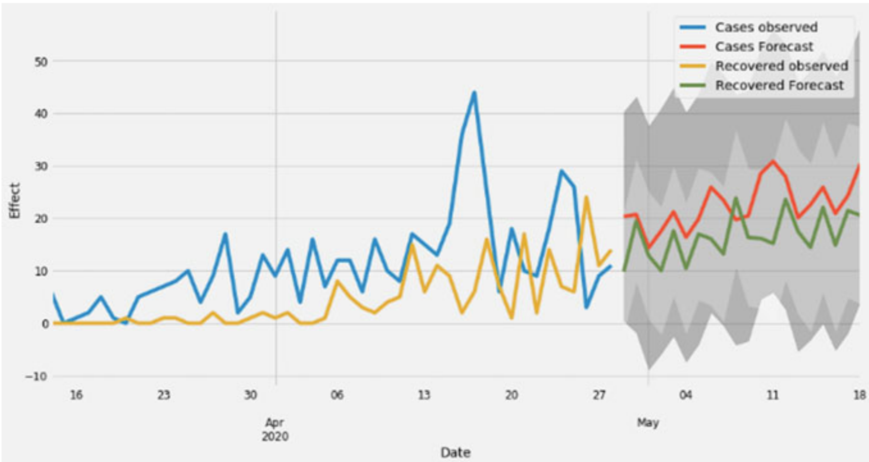


Fig. 13 Prediction of number of cases and recovered in Karnataka State

less than the predicted cases so, there might a possibility that the cases may become very high in coming days.

In Fig. 14, the predicted number of cases and recovered number of the Kerala state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is more than the predicted cases so, there might a possibility that the cases may become very less in coming days.

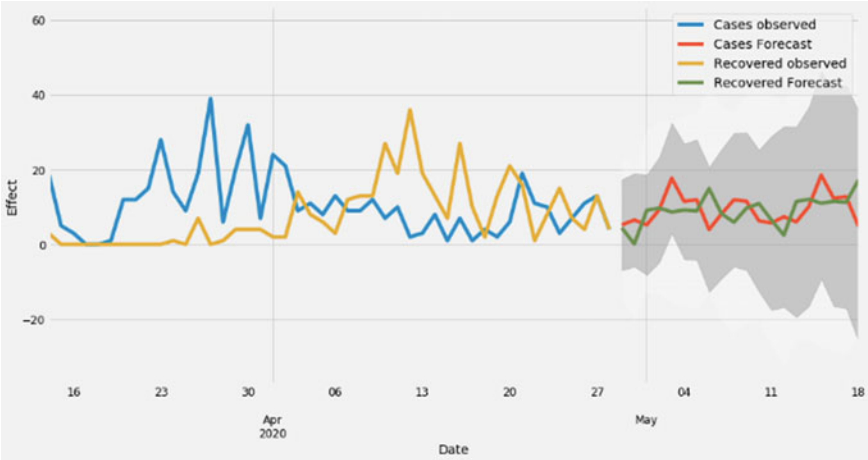


Fig. 14 Prediction of number of cases and recovered in Kerala State

In Fig. 15, the predicted number of cases and recovered number of the West Bengal state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is more than the predicted cases so, there might a possibility that the cases may become very less in coming days.

In Fig. 16, the predicted number of cases and recovered number of the Odisha state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about

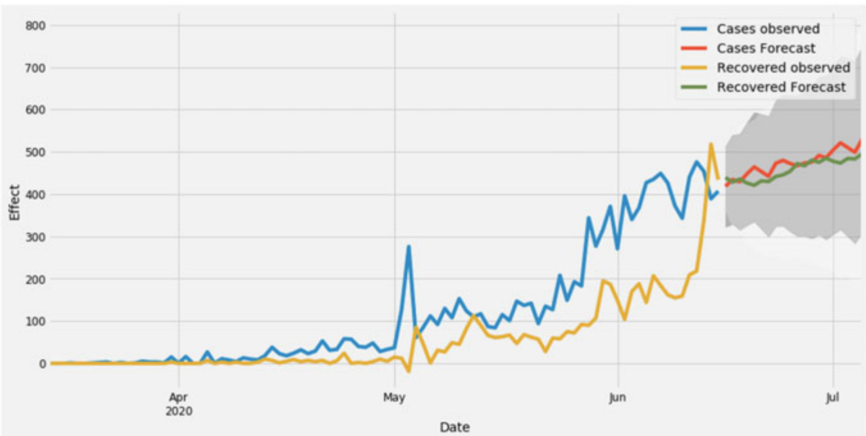


Fig. 15 Prediction of number of cases and recovered in West Bengal State

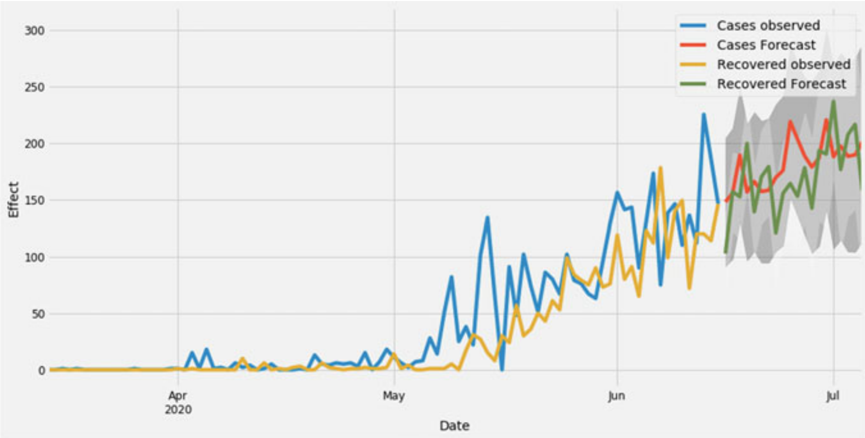


Fig. 16 Prediction of number of cases and recovered in Odisha State

the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is more than the predicted cases so, there might a possibility that the cases may become very less in coming days.

In Fig. 17, the predicted number of cases and recovered number of the Delhi state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is

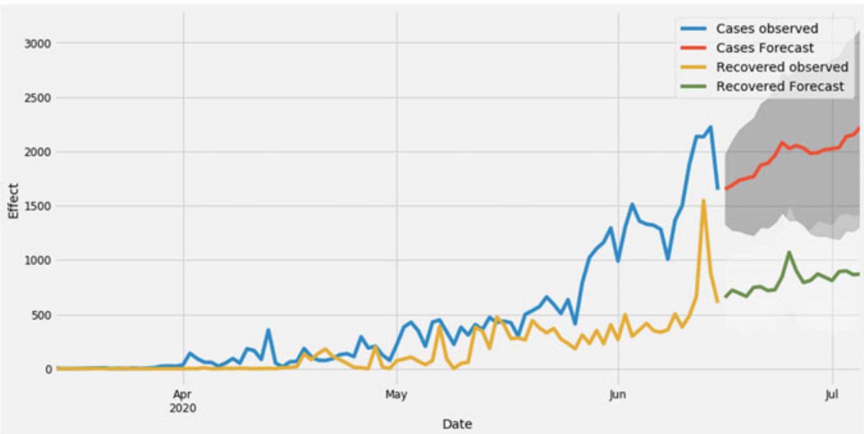


Fig. 17 Prediction of number of cases and recovered in Delhi State

less than the predicted cases so, there might a possibility that the cases may become very high in coming days.

In Fig. 18, the predicted number of cases and recovered number of the Bihar state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is more than the predicted cases so, there might a possibility that the cases may become very less in coming days.

In Fig. 19, the predicted number of cases and recovered number of the Madhya Pradesh state has been displayed using the line plot. In which the blue line gives the information about the present cases, red line tells the forecasted cases, yellow line tells about the number of recovered patients and green line tells about the forecasted number of recoveries. In the figure it is observed that the predicted number of recoveries is more than the predicted cases so, there might a possibility that the cases may become very less in coming days.

After the prediction of graphs, the data of the prediction produced by the ARIMA model is taken and interpreted to produce the estimated date from which the number of cases will reduce gradually is displayed in the Table 3. The table consists of some states on which the model was trained by the respective state time series data along with their estimated date of recovery. The date which was shown in the model tells about the day from which COVID-19 cases will reduce gradually in that state.

In Table 3, all the other states' data predictions have given RMSE values in the range of 4–9 but Bihar, Maharashtra, and West Bengal has shown RMSE values in the range of 20–30. The RMSE value are very much important in predicting the values, because if the prediction is done with greater RMSE value then lesser will be the accuracy of the value and if predicted value is with lesser RMSE value then

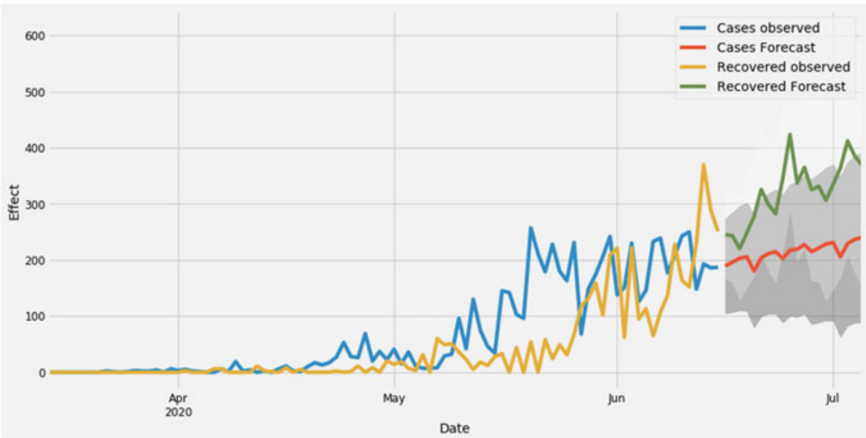


Fig. 18 Prediction of number of cases and recovered in Bihar State

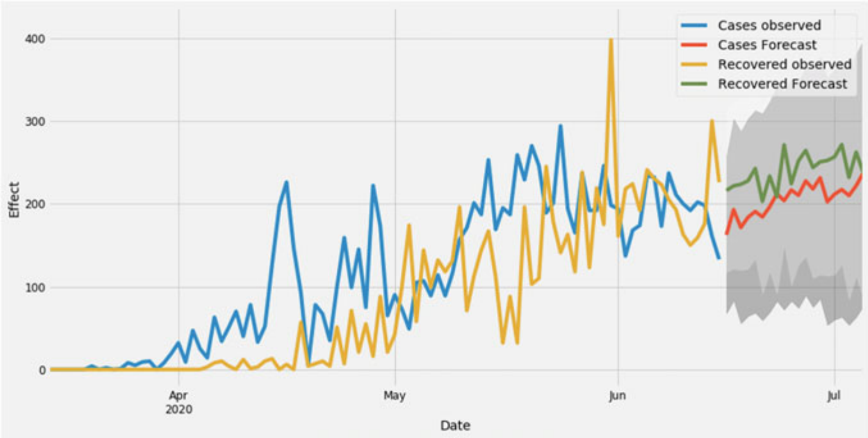


Fig. 19 Prediction of number of cases and recovered in Madhya Pradesh State

Table 3 Predicted date when the number of cases will reduce of various states is displayed

States of India	Recovery date (predicted)
Andhra Pradesh	28-07-2020
Assam	25-07-2020
Bihar	31-07-2020
Delhi	19-08-2020
Gujarat	20-07-2020
Haryana	21-08-2020
Karnataka	12-07-2020
Kerala	27-07-2020
Madhya Pradesh	30-07-2020
Maharashtra	21-08-2020
Jharkhand	31-07-2020
Jammu and Kashmir	25-07-2020
Ladakh	11-07-2020
Himachal Pradesh	06-07-2020
Goa	20-07-2020
Odisha	29-07-2020
Punjab	24-08-2020
Rajasthan	09-08-2020
Tamil Nadu	31-07-2020
Telangana	21-07-2020
Uttar Pradesh	13-09-2020
West Bengal	18-08-2020

greater will be the accuracy. So, the predicted information of those states is not as accurate and can change.

In Table 4, the details about the number of cases in each state along with the total number of cases throughout India is displayed. From the table, it is clear that the number of cases will increase up to 500 thousand from now till the 25-Sept-2020. So, the total number of cases throughout India may reach 10.5 million in the month of September 2020. The accuracy of the proposed model was predicted using the produced RMSE values. For RMSE value which were less than 5 the accuracy was 95% and above, for the values less 10 it was 90% and above. But those states with had a high RMSE value we had to consider the accuracy to be 80%. Therefore, from all the values of RMSE we can estimate that the accuracy of the model can be 90%.

Table 4 Predicted total number of cases till 23-09-2020

States of India	Total cases (predicted)
Andhra Pradesh	15,792
Assam	58,686
Bihar	61,679
Delhi	130,832
Gujarat	53,959
Haryana	106,447
Karnataka	88,186
Kerala	8,946
Madhya Pradesh	33,797
Maharashtra	363,422
Jammu and Kashmir	56,595
Jharkhand	18,690
Ladakh	6,256
Himachal Pradesh	4,565
Goa	8,989
Odisha	29,304
Punjab	7,878
Rajasthan	44,121
Tamil Nadu	160,605
Telangana	36,022
Uttar Pradesh	63,900
West Bengal	63,683
Total	1,422,354

6 Conclusion

The effect or impact of the pandemic COVID-19 virus in India was predicted using a statistical model ARIMA. The details about the ARIMA model was mentioned and the steps that were involved in the future prediction from the input data is also described. This paper also gives the graphical information about the COVID-19 effect on each state in India with 300 or more cases as per the records till 15-June-2020. As the accuracy of the model was 90% the predicted information may become true if the rules or precautions which were imposed by the government are strictly followed by the people of respective states, then the effect may reduce prior to the predicted dates and even the economic condition may come back to normal else there will be a rapid increase in the number of cases and the economy of India will collapse in the near future. And from the above mentioned dates, on average we can conclude that from August there will be slight downfall in the spreading of the disease when current conditions are imposed on the people by the respective state government and also the central government in India, if not the no. of cases may increase rapidly and can infect all the people throughout India.

References

1. https://api.covid19india.org/csv/
2. Perone, G.: An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3564865
3. Almasarweh, M., Alwadi, S.: ARIMA model in predicting banking stock market data. Mod. Appl. Sci. **12**, 309 (2018). https://doi.org/10.5539/mas.v12n11p309
4. Adebiyi, A., Adewumi, A., Ayo, C.: Stock price prediction using the ARIMA model. In: Proceedings—UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014 (2014). https://doi.org/10.1109/uksim.2014.67
5. Alsharif, M.H., Younes, M.K., Kim, J.: Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. Symmetry **11**, 240 (2019)
6. Awajan, A.M., Ismail, M.T., Al Wadi, S.: Improving forecasting accuracy for stock market data using EMD-HW bagging. PLoS ONE **13**(7), e0199582 (2018). https://doi.org/10.1371/journal.pone.0199582
7. Totakura, V., Devasekhar, V., Sake, M.: Prediction of stock trend for swing trades using long short-term memory neural network model. Int J Sci Tech Res **9**, 1918–1923 (2020)
8. Liu, C., Hoi, S.C.H., Zhao, P., Sun, J.: Online ARIMA algorithms for time series prediction. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16), pp. 1867–1873. AAAI Press (2016)
9. Weron, R.: Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int. J. Forecast. **30** (2014). https://doi.org/10.1016/j.ijforecast.2014.08.008
10. Taylor, S., Letham, B.: Forecasting at scale. Am. Stat. **72** (2017). https://doi.org/10.1080/00031305.2017.1380080
11. Krishna Sagar, B., Madhusudhana Reddy, E., Ramakrishna, S.: Network devices data visualization using Weka. Int. J. Pure Appl. Math. **118**(17), 599–608 (2018)
12. Armstrong, J.: Evaluating forecasting methods (2001). https://doi.org/10.1007/978-0-306-47630-3_20

13. Kumar, M., Anand, M.: An application of time series ARIMA forecasting model for predicting sugarcane production in India. Stud. Bus. Econ. **9**, 81–94 (2014)
14. Das, S., Ray, S., Sen, A., Siva, G., Das, S.: Statistical study on modeling and forecasting of jute production in West Bengal. Int. J. Curr. Microbiol. Appl. Sci. **8**, 1719–1730 (2019). https://doi.org/10.20546/ijcmas.2019.807.204
15. Petrevska, B.: Predicting tourism demand by A.R.I.M.A. models. Econ. Res.-Ekonomska Istraživanja **30**, 939–950 (2017). https://doi.org/10.1080/1331677X.2017.1314822
16. Ray, S.: Statistical modeling and forecasting of ARIMA and ARIMAX models for food grains production and net availability of India. **8**, 296–309 (2020). https://doi.org/10.18006/2020.8(3).296.309
17. Gavirangaswamy, V., Gupta, G., Gupta, A., Agrawal, R.: Assessment of ARIMA-based prediction techniques for road-traffic volume. In: Proceedings of the 5th International Conference on Management of Emergent Digital EcoSystems, MEDES 2013, pp. 246–251 (2013). https://doi.org/10.1145/2536146.2536176
18. Yadav, R.K., Balakrishnan, M.: Comparative evaluation of ARIMA and ANFIS for modeling of wireless network traffic time series. J. Wirel. Commun. Netw. **2014**, 15 (2014). https://doi.org/10.1186/1687-1499-2014-15
19. Duarte, D., Faerman, J.: Comparison of time series prediction of healthcare emergency department indicators with ARIMA and prophet, pp. 123–133 (2019). https://doi.org/10.5121/csit.2019.91810
20. Suarez, C., Salcedo, O., Diaz, A.: An ARIMA model for forecasting Wi-Fi data network traffic values. Ing. Invest. **29**, 65–69 (2009)
21. Villacorta Cardoso, C., Cruz, G.: Forecasting natural gas consumption using ARIMA models and artificial neural networks. IEEE Latin Am. Trans. **14**, 2233–2238 (2016). https://doi.org/10.1109/TLA.2016.7530418
22. Nugroho, A., Simanjuntak, B.H.: ARMA (Autoregressive Moving Average) model for prediction of rainfall in regency of Semarang-Central Java-Republic of Indonesia (2014)
23. Sidiq, M.: Forecasting rainfall with time series model. IOP Conf. Ser.: Mater. Sci. Eng. **407**, (2018). https://doi.org/10.1088/1757-899X/407/1/012154

Mr. Varun Totakura is currently working as a Software Developer in Tata Consultancy Services, Hyderabad, Telangana, India. He has completed his Graduation in Bachelor of Technology in Computer Science & Engineering from Guru Nanak Institutions Technical Campus affiliated to Jawaharlal Nehru Technological University, Hyderabad. He is a passionate Developer and also a Research enthusiast who is interested in the field of Artificial Neural Networks, Computer Vision, Data Analysis, Deep Learning. He has developed lot of projects and also published 10+ research papers on Machine Learning.

Dr. E. Madhusudhana Reddy is currently working as a professor in the department of Computer Science at Guru Nanak Institutions Technical Campus affiliated to Jawaharlal Nehru Technological University Hyderabad. He has Ph.D. in Computer Science from Sri Venkateswara University, Tirupati and M. Tech. from Punjabi University, Patiala. He is a research enthusiast who works in the field of Network Security, Computer Vision, Data Analysis, Deep Learning. He also published over 60 papers and has one patent and have authored 4 books or monographs. He has 16+ years or Teaching experience in the field of Computer Science. He guides the Ph.D. Scholars from various universities like JNTUH, Kakinada and many more.