Lecture Notes in Electrical Engineering 980

Md. Sazzad Hossain
Satya Prasad Majumder
Nazmul Siddique
Md. Shahadat Hossain *Editors*

The Fourth Industrial Revolution and Beyond

Select Proceedings of IC4IR+

Springer

Lecture Notes in Electrical Engineering

Volume 980

Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy
Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico
Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India
Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany
Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China
Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany
Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China
Gianluigi Ferrari, Università di Parma, Parma, Italy
Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain
Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany
Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA
Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt
Torsten Kroeger, Stanford University, Stanford, CA, USA
Yong Li, Hunan University, Changsha, Hunan, China
Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA
Ferran Martín, Departament d’Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore
Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany
Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA
Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany
Subhas Mukhopadhyay, School of Engineering and Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand
Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA
Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan
Luca Oneto, Department of Informatics, BioEngineering, Robotics and Systems Engineering, University of Genova, Genova, Genova, Italy
Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi “Roma Tre”, Rome, Italy
Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Gan Woon Seng, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany
Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal
Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China
Walter Zamboni, DIEM—Università degli studi di Salerno, Fisciano, Salerno, Italy
Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Md. Sazzad Hossain · Satya Prasad Majumder ·
Nazmul Siddique · Md. Shahadat Hossain
Editors

The Fourth Industrial Revolution and Beyond

Select Proceedings of IC4IR+

Springer

Editors
Md. Sazzad Hossain
Department of Computer Science
and Engineering
University Grants Commission of Banglade
Dhaka, Bangladesh

Satya Prasad Majumder
Department of Electrical and Electronic
Engineering
Bangladesh University of Engineering
and Technology
Dhaka, Bangladesh

Nazmul Siddique
School of Computing, Engineering,
and Intelligent Systems
University of Ulster
Londonderry, UK

Md. Shahadat Hossain
Department of Computer Science
and Engineering
University of Chittagong
Chittagong, Bangladesh

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-19-8031-2 ISBN 978-981-19-8032-9 (eBook)
https://doi.org/10.1007/978-981-19-8032-9

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Organization

Chief Patron

Prof. Datuk Ts. Dr. Wahid Bin Razzaly, Vice-chancellor, UTHM, Malaysia

Steering Committee

Kazi Shahidullah, Chairman, UGC of Bangladesh
Dil Afroza Begum, Member, UGC of Bangladesh
Md. Sazzad Hossain, Member, UGC of Bangladesh
Muhammad Alamgir, Member, UGC of Bangladesh
Biswajit Chanda, Member, UGC of Bangladesh
Md. Abu Taher, Member, UGC of Bangladesh

Local Advisory Committee

Dil Afroza Begum, Member, UGC
A. B. M. Siddique Hossain, American International University-Bangladesh
A. Z. M. Touhidul Islam, Rajshahi University
Mohammad Kaykobad, Brac University
M. Rizwan Khan, United International University
M. M. A. Hashem, Khulna University of Engineering and Technology
M. Lutfar Rahman, Daffodil International University
Mohammod Abul Kashem, Dhaka University of Engineering and Technology
Mozammel Huq Azad Khan, East West University
Nasim Akhtar, Chandpur University of Science and Technology
S. I. Khan, Brac University

Subrata Kumar Aditya, Sheikh Hasina University
Syed Akhter Hossain, University of Liberal Arts Bangladesh
Suraiya Pervin, University of Dhaka

International Advisory Committee

Muhammad Alamgir, Member, University Grants Commission
Alamgir Hossain, CamTech, Cambodia
Anirban Bandyopadhyay, NIMS, Japan
Ayman Alfalou, University of Brittany, France
Ekram Hossain, University of Manitoba, Canada
Latifur Khan, University of Texas, USA
M. A. Karim, University of Massachusetts, USA
M. Ashraful Alam, Purdue University, USA
Marek Perkowski, Portland State University, Oregon, USA
Mohammad S. Alam, Texas A&M University, USA
Mustafizur Rahman, Intel Co., USA
M. Zahid Hasan, Princeton University, USA
N. Kishi, University of Electro-communication, Japan
Ramjee Prasad, Aarhus University, Denmark
Ranjan Gangopadhy, LNMIIT, India
Vijayan Asari, University of Dayton, USA
Saifur Rahman, Virginia Tech, USA
Syed Ishtiaque Ahmed, University of Toronto, Canada
Taher Saif, Stanford University, USA
Kanad Ray, Amity University, India

Organizing Committee Chair

Md. Sazzad Hossain, Member, University Grants Commission

Organizing Committee

Md. Sarwar Morshed, Ahsanullah University of Science and Technology
M Mahfuzul Islam, Bangladesh University of Engineering and Technology
A K M Zakir Hossain, Bangladesh Agricultural University
Mamunur Rashid, University of Dhaka
Md. Shah Azam, Rajshahi University
Mohammad Tofayal Ahmed, Jashore University of Science and Technology

Md. Roshidul Hasan, Bangabandhu Sheikh Mujibur Rahman Agricultural University
Shamim Al Mamun, Jahangirnagar University
Ferdous Zaman, University Grants Commission
Sultan Mahmud Bhuiyan, University Grants Commission
A. K. M. Shamsul Arefin, University Grants Commission
Md. Kamal Hossain, University Grants Commission
Md. Fakhrul Islam, University Grants Commission
Md. Omar Faruque, University Grants Commission
Durga Rani Sarker, University Grants Commission
Md. Shah Alam, University Grants Commission
Muhammad Jaminur Rahman, University Grants Commission
M. Maksudur Rahman Bhuiyan, University Grants Commission
Md. Rezaul Karim Hawlader, University Grants Commission
Nahid Sultana, University Grants Commission
Bishnu Mallick, University Grants Commission
Probir Chandra Das, University Grants Commission

Technical Program Committee Chair

Satya Prasad Majumder, VC, Bangladesh University of Engineering and Technology

Technical Program Committee

Md. Shahadat Hossain, University of Chittagong
M. Shamim Kaiser, Jahangirnagar University
Md. Abdur Razzaque, University of Dhaka
Mufti Mahmud, Nottingham Trent University, UK
Nazrul Islam, Military Institute of Science and Technology
Md. Atiqur Rahman Ahad, University of Dhaka
Nazmul Siddique, Ulster University, UK
Mohammad Shorif Uddin, Jahangirnagar University
Celia Sahanaz, Bangladesh University of Engineering and Technology

Track Committee

Abdullah-Al Nahid, Khulna University, BD
Ahmed Wasif Reza, East West University, BD
Saadia Alam, University of Hyogo, Japan
Fazlul Hasan Siddiqui, Dhaka University of Engineering and Technology, BD

M. M. A. Hasehm, Khulna University of Engineering and Technology, BD
Md. Obaidur Rahman, Dhaka University of Engineering and Technology, BD
Md. Abu Layek, Jagannath University, BD
Md. Saiful Islam, Bangladesh University of Engineering and Technology, BD
Muhammad Golam Kibria, University of Liberal Arts Bangladesh, BD
Farhana Sarker, University of Liberal Arts Bangladesh, BD
Fida Hasan, RMIT University, BD
Hussain Nyeem, Military Institute of Science and Technology, BD
M. Shahidur Rahman, Shahjalal University of Science and Technology, BD
M. M. A. Hashem, Military Institute of Science and TechnologyET, BD
Md. Atiqur Rahman Ahad, University of Dhaka, BD
Md. Golam Mostafa, Military Institute of Science and Technology, BD
Md. Hossam-E Haider, Military Institute of Science and Technology, BD
Md. Kafiul Islam, Independent University, Bangladesh, BD
Md. Sazzadur Rahman, IIT, Jahangirnagar University, BD
Md. Abdul Alim, Military Institute of Science and Technology, BD
Md. Abdur Rahman, American International University-Bangladesh
Md. Farhad Hossain, Bangladesh University of Engineering and Technology
Md. Golam Rashed, University of Rajshahi
Md. Saiful Islam, Bangladesh University of Engineering and Technology
Mohammad Abu Yousuf, Jahangirnagar University
Mohammad Nurul Huda, United International University
Md. Shahadat Hossain, University of Chittagong
Mohammad Shamsul Arefin, Chittagong University of Engineering and Technology
Mohammed Hoque, Chittagong University of Engineering and Technology
Mohammed Eunus Ali, Bangladesh University of Engineering and Technology
Mufti Mahmud, Nottingham Trent University, UK
Muhammad Nazrul Islam, Military Institute of Science and Technology
Nabeel Mohammed, North South University
Nasima Begum, University of Asia Pacific
A. K. M. Fazlul Haque, Daffodil International University
Shaikh Anowarul Fattah, Bangladesh University of Engineering and Technology
S. M. Abdur Razzak, Rajshahi University of Engineering and Technology
Rameswar Debnath, Khulna University
Rashed Mustafa, University of Chittagong
Risala Tahsin Khan, Jahangirnagar University
S. M. Riazul Islam, Sejong University, Korea
Sejutib Rahman, University of Dhaka
M. Shamim Kaiser, Jahangirnagar University
Shamim Al Mamun, Jahangirnagar University
Syed Galib, Jashore University of Science and Technology

Technical Review Committee

A. F. M. Shahab Uddin, KyungHee University
A. N. M. Shahebul Hasan, American International University-Bangladesh, BD
A. S. M. Touhidul Hasan, University of Asia PacificD
A. B. M. Alim Al Islam, Bangladesh University of Engineering and Technology, BD
A. B. M. Aowlad Hossain, Khulna University of Engineering and Technology, BD
A. H. M. Hoque, University of Chittagong, BD
A. K. M. Fazlul Haque, Daffodil International University, BD
A. S. M. Sanwar Hosen, JNU, South Korea
Abdullah-Al Nahid, Khulna University, BD
Abir Ahmed, American International University-Bangladesh, BD
A. B. M. Rashid, Bangladesh University of Engineering and Technology, BD
Abu Hena Shatil, American International University-Bangladesh, BD
Abu Nowshed Chy, University of Chittagong, BD
Abu Shamim Mohammad Arif, Khulna University, BD
Abul Kalam Azad, University of Liberal Arts Bangladesh, BD
Afsana Begum, Daffodil International University, BD
Ahsanur Rahman, NSU, BD
A. K. M. Baki, Ahsanullah University of Science and Technology, BD
Aminul Islam, Military Institute of Science and Technology, BD
Aminul Islam, Khulna University, BD
Anas Abdullah Alvi, American International University-Bangladesh, BD
Anindya Iqbal, Bangladesh University of Engineering and Technology, BD
Anupam Bairagi, Khulna University, BD
Arif Ahmad, Shahjalal University of Science and Technology, BD
Ashikur Rahman, Bangladesh University of Engineering and Technology, BD
Asif Mahfuz, American International University-Bangladesh, BD
Atanu Shome, Khulna University, BD
Alessandra Pedrocchi, Politechnico di Milano, Italy
Alessandro Gozzi, IIT, Italy
Amir Hussain, ENU, UK
Anirban-Bandyopadhyay, NIMS, Japan
Antesar Shabut, LTU, UK
Atik Mahabub, Concordia University, Canada
Bishwajit Banik Pathik, American International University-Bangladesh, BD
Bobby Barua, Ahsanullah University of Science and Technology, BD
Belayat Hossain, LU, UK
Bernd Kuhn, OIST, Japan
Bo Song, USQ, Australia
Chinmay Bepery, Patuakhali Science and Technology University, BD
Cosimo Ieracitano, UNIRC, Italy
Cris Calude, University of Auckland, New Zealand
Daniel K. Wójcik, Nencki Institute of Experimental Biology, Poland

Daniel Marcus, UW, USA
David Brown, Nottingham Trent University, UK
Davide Zoccolan, SISSA, Italy
Derong Liu, UIC, USA
Dipankar Das, University of Rajshahi, BD
Effat Jahan, American International University-Bangladesh, BD
Egidio D'Angelo, University of Pavia, Italy
Eleni Vasilaki, The University of Sheffield, UK
Francesco Carlo Morabito, UNIRC, Italy
Francesco Papaleo, University of Padova and IIT, Italy
Fahim Alam, University of Chittagong, BD
Farhadur Arifin, American International University-Bangladesh, BD
Farhana Sarker, University of Liberal Arts Bangladesh, BD
Farhana Tazmim Pinki, Khulna University, BD
Fazlul Hasan Siddiqui, Dhaka University of Engineering and Technology, BD
Ferdous Shaun, American International University-Bangladesh, BD
Feroz Ahmed, Independent University, Bangladesh
Fida Hasan, RMIT University
Gabriella Panuccio, IIT, Italy
Gaute Einevoll, NMBU, Norway
Giacomo Indiveri, UZH, Switzerland
Giancarlo Ferregno, Politechnico di Milano, Italy
Giorgio A. Ascoli, GMU, USA
Golam Dastoger Bashar, Boise State University, USA
Guenther Zeck, NMI, Germany
Gustavo Deco, UPF, Spain
Hussain Nyeem, Military Institute of Science and Technology, BD
Imtiaz Mahmud, Kyungpook National University
Iqbal Sarker, Chittagong University of Engineering and Technology, BD
Iqbal Aziz Khan, Rajshahi University, BD
Jabed Faysal, Military Institute of Science and Technology, BD
Joyprokash Chakrabartty, Chittagong University of Engineering and Technology, BD
Imtiaz Mahmud, KNU, Korea
Joarder Kamruzzaman, Federation University, Australia
Jonathan Mappelli, University of Modena, Italy
Kafi Rahman, Truman State University, BD
Kamruddin Nur, American International University-Bangladesh, BD
Kawshik Shikder, American International University-Bangladesh, BD
Kazi Firoz Ahmed, American International University-Bangladesh, BD
Kazi Masudul Alam, Military Institute of Science and Technology, BD
Kazi Md. Rokibul Alam, Khulna University of Engineering and Technology, BD
Kazi Sakib, University of Dhaka, BD
Khaleda Ali, Dhaka University, BD
Khandaker Tabin Hasan, American International University-Bangladesh, BD

K. C. Santosh, USD, USA
Lu Cao, Saitama University, Japan
Luca Benini, ETH, Switzerland
Luca Berdondini, IIT, Italy
Luciano Gamberini, University of Padova, Italy
Lafifa Jamal, University of Dhaka, BD
Lasker Ershad Ali, Khulna University, BD
Lutfa Akter, Bangladesh University of Engineering and Technology
M. Akhtaruzzaman, Military Institute of Science and Technology, BD
M. Kabiruzzaman, American International University-Bangladesh, BD
M. M. Manjurul Islam, American International University-Bangladesh, BD
M Raihan, North Western University, Khulna, BD
M. A. H. Akhand, Khulna University of Engineering and Technology, BD
M. Shahidur Rahman, Shahjalal University of Science and Technology, BD
M. M. A. Hashem, Khulna University of Engineering and Technology, BD
Mahfuzul Hoq Chowdhury, Chittagong University of Engineering and Technology, BD
Manan Noor, Jahangirnagar University, BD
Manishankar Mondal, Khulna University, BD
Manowarul Islam, JNU, BD
Marium-E –Jannat, Shahjalal University of Science and Technology, BD
Md. Abdul Masud, Patuakhali Science and Tech. University, BD
Md. Aktaruzzaman, Islamic University, BD
Md. Arefin, Bangladesh University of Business and Technology, BD
Md. Al Mamun, Rajshahi University of Engineering and Technology, BD
Md. Ashif Islam Oni, American International University-Bangladesh, BD
Md. Atiqur Rahman Ahad, University of Dhaka, BD
Md. Awal, Khulna University, BD
Md. Dulal Haque, Hajee Mohammad Danesh Science and Technology University, BD
Md. Ezharul Islam, Jahangirnagar University, BD
Md. Forhad Rabbi, Shahjalal University of Science and Technology, BD
Md. Foysal Haque, Dong-A University
Md. Halim, Military Institute of Science and Technology, BD
Md. Islam, University of Rajshahi, BD
Md. Jaber Nahian, BUP, BD
Md. Mahadi Hasan Nahid, Shahjalal University of Science and Technology, BD
Md. Mehedi Hasan, University of Dhaka, BD
Md. Mostofa Kamal Rasel, East West University, BD
Md. Nasim Adnan, Jashore University of Science and Technology, BD
Md. Nazmul Hasan, University of Ulsan
Md. Obaidur Rahman, Dhaka University of Engineering and Technology, BD
Md. Rahman, UITS, BD
Md. Rajib Hossain, Chittagong University of Engineering and Technology, BD
Md. Rakibul Islam, University of Wisconsin - Eau Claire

Md. Rifat Ahmmd, Rashid University of Liberal Arts Bangladesh, BD
Md. Samsuzzaman, Patuakhali Science and Technology University, BD
Md. Sazzadur Rahman, Jahangirnagar University, BD
Md. Shahidul Islam, Jahangirnagar University, BD
Md. Tanzil Shahria, University of Wisconsin-Milwaukee
Md. Zahidul Islam, Islamic University
Md. Zulfiker Mahmud, Jagannath University, BD
Md. Islam, University of Dhaka, BD
Md. Abdul Alim, Khulna University, BD
Md. Abu Yousuf, Jahangirnagar University, BD
Md. Abu Layek, Jagannath University, BD
Md. Ahsan Habib, Mawlana Bhashani Science and Technology University, BD
Md. Alamgir Hossain, Islamic University, BD
Md. Anwarul Abedin, Dhaka University of Engineering and Technology, BD
Md. Arafat Hossain, Khulna University of Engineering and Technology, BD
Md. Ashraful Islam, Rajshahi University, BD
Md. Delowar Hossain, Hajee Mohammad Danesh Science and Technology University, BD
Md. Eamin Rahmn, Shahjalal University of Science and Technology, BD
Md. Farhad Hossain, Bangladesh University of Engineering and Technology, BD
Md. Faruque Hossain, Khulna University of Engineering and Technology, BD
Md. Foisal Hossain, Khulna University of Engineering and Technology, BD
Md. Forkan Uddin, Bangladesh University of Engineering and Technology, BD
Md. Golam Rashed, Rajshahi University, BD
Md. Habibur Rahman, BDU
Md. Hasanul Kabir, IUT, BD
Md. Hossain, Bangladesh University of Business and Technology, BD
Md. Humayun Kabir, American International University-Bangladesh
Md. Humayun Kabir, Jahangirnagar University, BD
Md. Iqbal Bahar Chowdhury, United International University, BD
Md. Jahedul Islam, Khulna University of Engineering and Technology, BD
Md. Kafiul Islam, Independent University, Bangladesh
Md. Kamrul Hassan, American International University-Bangladesh
Md. Liakot Ali, Bangladesh University of Engineering and Technology, BD
Md. Mahabub Hossain, Hajee Mohammad Danesh Science and Technology University, BD
Md. Mahbub Hasan, Khulna University of Engineering and Technology, BD
Md. Mahbubur Rahman, Military Institute of Science and Technology, BD
Md. Meganur Rhaman, Ahsanullah University of Science and Technology, BD
Md. Mizanur Rahman, Military Institute of Science and Technology, BD
Md. Motaharul Islam, United International University, BD
Md. Motiur Rahman, Mawlana Bhashani Science and Technology University, BD
Md. Musfique Anwar, Jahangirnagar University, BD
Md. Nahiyan Uddin, University of Asia Pacific, BD
Md. Nawab Yousuf Ali, East West University, BD

Md. Rabiul Islam, American International University-Bangladesh
Md. Rafiqul Islam, Khulna University of Engineering and Technology, BD
Md. Ragib Amin Nihal, University of Dhaka, BD
Md. Rifat Hazari, American International University-Bangladesh, BD
Md. Saddam Hossain Mukta, United International University, BD
Md. Saiful Islam, Bangladesh University of Engineering and Technology
Md. Salah Uddin Yusuf, Khulna University of Engineering and Technology, BD
Md. Samsuzzaman Sobuz, Patuakhali Science and Technology University, BD
Md. Sanaullah Chowdhury, University of Chittagong, BD
Md. Saniat Rahman Zishan, American International University-Bangladesh, BD
Md. Shahariar Parvez, American International University-Bangladesh, BD
Md. Shahjahan, Khulna University of Engineering and Technology, BD
Md. Shahriar Mahbub, Ahsanullah University of Science and Technology, BD
Md. Shamim Ahsan, Military Institute of Science and Technology, BD
Md. Sheikh Sadi, Khulna University of Engineering and Technology
Md. Sohel Mahmud Sher, Military Institute of Science and Technology, BD
Md. Tanvir Hasan, Jashore University of Science and Technology, BD
Md. Tariq Hasan, Military Institute of Science and Technology, BD
Mehedi Hasan, American International University-Bangladesh, BD
Milon Biswas, Bangladesh University of Business and Technology
Mir Md. Jahangir Kabir, Rajshahi University of Engineering and Technology, BD
M. M. Shahriar Maswood, Khulna University of Engineering and Technology, BD
Mohammad Shidujaman, American International University-Bangladesh
M. Abdullah Al Mumin, Shahjalal University of Science and Technology, BD
Mohammad Abu Yousuf, Jahangirnagar University, BD
Mohammad Farhad Bulbul, Jashore University of Science and Technology, BD
Mohammad Khairul Islam, University of Chittagong, BD
Mohammad Mamun Elahi, United International University, BD
Mohammad Nasir Uddin, American International University-Bangladesh
Mohammad Nurul Huda, United International University, BD
M. Rashedur Rahman, NSU, BD
M. Shamsul Arefin, Chittagong University of Engineering and Technology, BD
M. Moshiul Hoque, Chittagong University of Engineering and Technology, BD
Mohidul Islam, Military Institute of Science and Technology, BD
Molla Hussein, University of Asia Pacific, BD
Monir Hossen, Khulna University of Engineering and Technology, BD
Monir Morshed, Mawlana Bhashani Science and Technology University, BD
Mosabber Uddin Ahmed, University of Dhaka, BD
Mostafa Zaman Chowdhury, Khulna University of Engineering and Technology, BD
Mostofa Nasir, Mawlana Bhashani Science and Technology University, BD
Muammer Arif, Military Institute of Science and Technology, BD
Mufti Mahmud, Nottingham Trent University, UK
Muhammad Abdullah Adnan, Bangladesh University of Engineering and Technology, BD

Muhammad Asad Rahman, Chittagong University of Engineering and Technology, BD
Muhammad Golam Kibria, University of Liberal Arts Bangladesh, BD
Muhammad Hasan, University of Liberal Arts Bangladesh, BD
Muhammad Nazrul Islam, Military Institute of Science and Technology, BD
Muhammad Sajjadur Rahim, Rajshahi University, BD
Muhammad Shahin Uddin, Mawlana Bhashani Science and Technology University, BD
Muhammed Patwary, International Islamic University Chittagong, BD
Mustafa Chowdhury, Independent University, Bangladesh
Nabeel Mohammed, North South University
M. Ali Akber Dewan, Athabasca University, Canada
Manjunath Aradhya, JSS STU, India
Manohar Das, Oakland University, USA
Marco Mongillo, University of Padova, Italy
Martin McGinnity, Ulster University, UK
Marzia Hoque-Tania, Oxford University, UK
Mathew Diamond, SISSA, Italy
Nadia Nowshin, American International University-Bangladesh, BD
Nafees Mansoor, University of Liberal Arts Bangladesh, BD
Nafisa Noor, NSU, BD
Nafiz Ahmed Chisty, American International University-Bangladesh, BD
Nafiz Imtiaz Khan, Military Institute of Science and Technology, BD
Naruttam Kumar Roy, Khulna University of Engineering and Technology, BD
Nasrin Akhter, Khulna University of Engineering and Technology, BD
Nasrin Sultana, Hajee Mohammad Danesh Science and Technology University, BD
Nihad Karim Chowdhury, University of Chittagong, BD
Niloy Sikder, Khulna University, BD
Nirjhor Tahmidur Rouf, American International University-Bangladesh, BD
Norhidayah Azman, Universiti Sains Islam Malaysia
Nowshin Alam, American International University-Bangladesh, BD
Nusrat Sharmin, Military Institute of Science and Technology, BD
Nelishia Pillay, UP, South Africa
Nikola Kasabov, AUT, New Zealand
Nilanjan Dey, JIS University, India
Ning Zhong, Maebashi Institute of Technology, Japan
Noushath Shaffi, CAS, Oman
Ofer Yizhar, Weizmann Institute of Science, Israel
Omprakash Kaiwartya, Nottingham Trent University, UK
Omar Sharif, Chittagong University of Engineering and Technology, BD
Pallab Kumar Choudhury, Khulna University of Engineering and Technology, BD
Pintu Shill, Khulna University of Engineering and Technology, BD
Pranab Kumar Dhar, Chittagong University of Engineering and Technology, BD
Paolo Del Giudice, ISS, Italy
Paolo Massobrio, UniGe, Italy

Patrick Ruther, University of Freiburg, Germany
Rafia Nishat Toma, University of Ulsan
Raja Hasan, American International University-Bangladesh, BD
Rameswar Debnath, Military Institute of Science and Technology, BD
Rashed Mazumder, Jahangirnagar University, BD
Rashed Mustafa, University of Chittagong, BD
Riasat Khan, NSU, BD
Rifat Shahriyar, Bangladesh University of Engineering and Technology, BD
Risala Tahsin Khan, Jahangirnagar University, BD
Ruhul Amin, Fordham University
S. M. Imrat Rahman, American International University-Bangladesh
Sajeeb Saha, JnU, BD
Sajjad Waheed, Mawlana Bhashani Science and Technology University, BD
Shahriar Rahman, United International University, BD
Shamim Ahmad, Rajshahi University, BD
Shamim Al Mamun, Jahangirnagar University, BD
Shazzad Hosain, NSU, BD
Sifat Momen, NSU, BD
Silvia Ahmed, NSU, BD
Sohrab Nizami, University of Queensland
Subrata Aditya, Dhaka University, Bangladesh
Sudipta Kar, Amazon
Sujan Sarker, University of Dhaka
Syed Galib, Jashore University of Science and Technology, BD
S. M. Riazul Islam, Sejong University, South Korea
Tahseen Asma Meem, American International University-Bangladesh
Tama Fouzder, University of Liberal Arts Bangladesh, BD
Tanvir Rahman, Mawlana Bhashani Science and Technology University, BD
Tapotosh Ghosh, NUB, BD
Taslima Ferdaus, Daffodil International University, BD
Tasmiah Anannya, Military Institute of Science and Technology, BD
Tawsin Uddin Ahmed, University of Chittagong, BD
Tofael Ahmed, Chittagong University of Engineering and Technology, BD
Uzzal Acharjee, Jagannath University, BD
Zinat Tasneem, Rajshahi University of Engineering and Technology, BD

Preface

The International Conference on 4th Industrial Revolution and Beyond, IC4IR 2021, is being organized by the University Grants Commission (UGC) of Bangladesh to be held through both physical and virtual online formats from 10 and 11 December 2021. On the 100th Birth Anniversary of the Father of the Nation of Bangladesh, Bangabandhu Sheikh Mujibur Rahman and 50th anniversary of independence of the country, the UGC is planning to create a premier international forum for bringing together researchers and practitioners from diverse domains to share cutting-edge research results obtained through the application of artificial intelligence, the internet of things, data analytics, and cloud computing to solve problems in the industrial domain. IC4IR 2021 also promotes the sharing and dissemination of innovative and practical development of methodologies and technologies with industrial applications through the exchange and dissemination of best practices. The scope of the conference includes Artificial Intelligence; Robotics and Automation; IoT and Smart Agriculture; Data Analytics and Cloud Computing; Communication and Networks Signal and Natural Language Processing. Researchers of home and abroad are requested to submit your research paper which can resolve our local problems, to improve life and the environment utilizing state-of-the-art 4IR techniques, and to help to achieve the sustainable development goal.

The IC4IR 2021 attracted 525 full papers from 16 countries in six tracks. These tracks include—Artificial Intelligence; Robotics and Automation; IoT and Smart Agriculture; Data Analytics and Cloud Computing; Signal and Natural Language Processing; and Communication and Networking. The submitted papers underwent a double-blind review process, soliciting expert opinion from at least two experts: at least two independent reviewers, the track co-chair, and the respective track chair. Following rigorous review reports from the reviewers and the track chairs, the technical program committee has selected 59 high-quality full papers from 09 countries that were accepted for presentation at the conference. Due to the COVID-19 pandemic, the Organizing Committee decided to host the event in hybrid mode. However, the research community reacted amazingly in this challenging time. More than 2000 people attended the events, which featured keynote speeches from three Nobel laureates and six distinguished professors.

The book series will be insightful and fascinating for those interested in learning about the tool and technique of 4IR that explores the dynamics of exponentially increasing knowledge in core and related fields. We are thankful to the authors who have made a significant contribution to the conference and have developed relevant research and literature in computation and cognitive engineering.

We would like to express our gratitude to the Organizing Committee and the Technical Committee members for their unconditional support, particularly the Chair, the Co-Chair and the Reviewers. Special thanks to the Chairman and Members of UGC of Bangladesh for their thorough support. IC4IR 2021 could not have taken place without the tremendous work of the team and the gracious assistance. We are grateful to Mr. Aninda Bose, Ms. Sharmila Mary Panner Selvam, and other term members of Springer Nature for their continuous support in coordinating this volume publication. We would also like to thank vice chancellor of the public and private universities for continuous support. Last but not the least, we thank all of our contributors and volunteers for their support during this challenging time to make IC4IR 2021 a success.

Dhaka, Bangladesh — Md. Sazzad Hossain
Dhaka, Bangladesh — Satya Prasad Majumder
Londonderry, UK — Nazmul Siddique
Chittagong, Bangladesh — Md. Shahadat Hossain
March 2022

Contents

Devices and Communication

Emerging Applications

Healthcare

Smart Signals and NLP Agriculture

About the Editors

Md. Sazzad Hossain is a distinguished professor, researcher, academic teaching advisor with extensive experience in the ICT sector, and writer. Currently, he is working as a full-time member of the University Grants Commission of Bangladesh. He is also a director of the Board of Directors of Bangladesh Communication Satellite Company Ltd (BCSCL). Dr. Sazzad is a masterly instructor in the ICT field with enormous knowledge of the subject matter. He joined as a professor as well as the head of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh in December 2009. He received his Ph.D. and master's degree in Electrical and Computer Engineering from Portland State University, Oregon, USA. He received his Bachelor of Science in Electrical System Network Engineering from Moscow Technical University (Popularly known as Moscow Power Institute), Moscow, Russia. He has twenty-six (26) years of experience in Information Technology. His research interest includes quantum computing, quantum algorithms, quantum entanglement, decoherence, natural computing, fault-tolerant computing, theoretical computer science, nanotechnology, low power VLSI design, reversible logic, software engineering, IoT, artificial intelligence, robotics, human-robot interaction. Based on his research findings, over 50 papers have been published in peer-reviewed international journals and conference proceedings.

Satya Prasad Majumder joined the position of vice chancellor of Bangladesh University of Engineering and Technology (BUET) in June 2020. He is a professor in the Department of Electrical and Electronic Engineering (EEE), BUET. He received his B.Sc. and M.Sc. in Electrical and Electronic Engineering from BUET, Bangladesh, in 1981 and 1985, respectively. In 1993, he obtained his Ph.D. degree in Optical Fiber Communication Engineering from the Indian Institute of Technology (IIT) Kharagpur, India. His research interests include optoelectronics and photonics, optical fiber communication systems, DWDM optical networks, soliton propagation, broadband wireless communications, mobile, and infrared networks. He supervised over 50 M.Sc. theses and 6 Ph.D. dissertations and published over 200 research papers in reputed refereed journals and conference proceedings.

Nazmul Siddique is with the School of Computing, Engineering, and Intelligent Systems, Ulster University. He obtained Dipl.-Ing. degree in Engineering Cybernetics and Automation Engineering from the TU Dresden, Germany, an M.Sc. in Computer Science from BUET, Bangladesh, and a Ph.D. in Intelligent Control from the Department of Automatic Control and Systems Engineering, University of Sheffield, England. He was awarded the prestigious Commonwealth Academic Staff Scholarship for Ph.D. at the University of Sheffield, England, and the German Government Scholarship for Dipl.-Ing at TU Dresden, Germany. His research interests include robotics, cybernetics, computational intelligence, nature-inspired computing, stochastic systems, and vehicular communication. He has published over 190+ research papers in the broad area of computational intelligence, vehicular communication, robotics, and cybernetics. He authored and co-authored five books published by John Wiley, Springer, and Taylor & Francis. He guest-edited eight special issues of reputed journals on Cybernetic Intelligence, Computational Intelligence, Neural Networks, and Robotics.

Md. Shahadat Hossain is a professor of Computer Science and Engineering at the University of Chittagong, Bangladesh since 2007. He is the first professor of Computer Science and Engineering in this region of Bangladesh. He did both his M.Phil. and Ph.D. in Computation from the University of Manchester Institute of Science and Technology (UMIST), the UK, in 1999 and 2002, respectively. He is continuing his research in other adventurous areas of computing such as health informatics, affective computing, e-government, smart cities, IoT, and the philosophy of computing. He has published 130 scholarly articles in reputed international journals and conferences. He is the author of several books. His jointly authored book entitled *Computing Reality*, published by Aoshima Research Institute (blue ocean press) in Tokyo, Japan, contributed significantly to enriching the knowledge of computer science and is indexed in the ACM Digital Library, published by the Association for Computing Machinery, as one of the important computing literature books.

AI for Detection and Prediction

Deep Learning-Based Rice Disease Prediction

Mahbuba Yesmin Turaba

Abstract Rice diseases are a prominent cause of reduced crop yield. Viruses, fungi, and bacteria might transmit all kinds of rice diseases. Early identification could have a significant impact in managing diseases and their spread. Automated rice disease identification systems could excel if they can have high predictive accuracy while requiring low computational resources. Nowadays, deep learning techniques are implemented in order to lessen human effort and time. In this paper, we present different experiments and findings related to the detection of three rice diseases: bacterial leaf blight, brown spot, and leaf smut. Multiple large pretrained models are fine-tuned on a relevant dataset. The fine-tuned models are then pruned and in one case quantized to achieve better performance. Most notably, the ResNet50-based model achieves 97% accuracy among all of them while requiring only 7200 FLOPS. In terms of correctness, this result is almost 4% better than previously published work while requiring a markedly lower number of FLOPS.

Keywords Disease prediction · Deep learning · Floating-point operation · Rice disease · Data augmentation

1 Introduction

Rice plant disease is one of the most dominant reasons for impacting a country's overall agricultural yield directly [1, 2]. Approximately, 10 to 15% of rice of total production is destroyed every year in Asia [3]. It is possible that rice plants could be affected by diseases at any time and bacteria, fungi, or viruses are mainly responsible for rice plant diseases. Common diseases include bacterial leaf blight, brown spot, leaf smut, leaf blast, and sheath blight [4].

Previously, image processing was remarkably popular for predicting and classifying infected plants. Diseases might vary for various plants in terms of color, size, or shape symptoms, causing difficulties for new farmers, as they have insufficient

M. Y. Turaba (✉)
North South University, Dhaka, Bangladesh
e-mail: mahbuba.turaba@Northsouth.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_1

knowledge about these conditions. In this case, constant monitoring might help in the prevention of rice plant diseases, but it is impossible for human beings to do so manually. In that case, different types of deep learning and machine learning techniques might be utilized to identify rice leaf diseases automatically, and this automatic system might lessen the required time in order to detect rice disease which might lead to higher crop productivity. It is important to point out that early detection is more effective for farmers to prevent diseases.

In our research, we have explored multiple pretrained models to detect three rice plant diseases including bacterial leaf blight, brown spot, and leaf smut and explored pruning and quantization to further optimize inference. The main contribution of our paper is summarized as below:

- Three different pretrained models VGG16, ResNet50, and Inception V3 are explored to predict rice disease, achieving 97% with ResNet50.
- Explored pruning and quantization to deliver a model that uses only 8200 flops. However, our best performing model only requires 7200 FLOPS.

Thus, these contributions could lead to autonomous rice disease prediction which helps to reduce crop destruction in the foreseeable future.

We have organized this paper into the following manners. We introduce a review of the literature related to rice disease prediction in Sect. 2. We discuss the main methodology of our approach in Sect. 3. Our experiments performed on rice disease datasets are presented in Sect. 4 in comparison with baseline methods. We discuss our result in Sect. 5. Lastly, we conclude and present our future work plans in Sect. 6.

2 Related Work

Rice is considered as staple food in many countries, and rice disease is responsible for catastrophic financial loss around the world [5]. Rice brown spot (BS) brings losses widely in the range of 4–52% [6]. Rice plant disease identification and prediction are recognized as a fascinating research area in which machine learning or deep learning could be applied for better performance. A review of relevant literature shows that the accuracy is much higher for recently explored deep learning-based approaches compared to other techniques.

Suman et al. [7] efficiently classified four types of diseases in rice using SVM classifier with 70% accuracy. But when it contrasts with others, it provides the lowest accuracy. On the other hand, Khaing et al. [8] used SVM for classification and identified four types of diseases leaf blast, leaf streak, BLB, and leaf brown spot and thus achieved 90% accuracy. But categorization of rice crop diseases was infeasible for SVM and PCA methodology . In addition, some researchers extract features based on color, shape, and texture [9]. However, Akila et al. [10] discovered R-FNN, R-CNN, and SSD to identify diseases and pests in complex environments and achieved 88% accuracy. Thus, this process seems time consuming. Also, Chowdhury et al.

[11] applied deep CNN-centered classification to detect diseases and perform 95% accuracy. But it requires more time to classify diseases as it is based on deep learning architecture.

In recent times, Fahrul et al. [12] showed reliability with eight kinds of diseases and 48 symptoms, but the performance seems not praiseworthy in comparison with the expert system. WeiJuan et al. [13] demonstrate that they could identify blast fungus, Magnaporthe grisea with 85% accuracy, but this method is not capable of finding other types of diseases. Moreover, Toran et al. [14] identified sheath blight, panicle blast, leaf blast, and brown spot with good recognition and reported 95.5% accuracy, but some diseases were not identified accurately.

Also, it is notable that Maohua et al. [15] demonstrate that disease could be identified in a faster and efficient way by means of PCA and neural network with 95.83% accuracy. But it could not recognize lesions perfectly.

In contrast, we propose various deep learning pretrained models to achieve higher prediction, and these are also efficient as it requires less computational resources and less cost.

3 Methods

In this paper, we propose different approaches to train our pretrained model and followed by pruning in order to improve the performance of transfer learning techniques. Our goal is to generate a new model that could beat existing works.

In order to perform rice disease prediction, the whole workflow is separated into four parts such as dataset, data augmentation, pretrained model selection training for rice disease prediction. We have used three pretrained models such as VGG16, ResNet50, and Inception V3 on the dataset and pruned the resulting models to make them more optimized for production.

3.1 Pretrained Architectures

Inception V3 Inception V3 is an updated version of the Inception series that includes various improvements [16]. It can help in object detection and image analysis [17]. Inception V3 has 48 convolutional layers. Factorized convolutional layers, dimension reduction, parallelization, and regularization are some of the techniques that are used in Inception V3. To compute computational efficiency and monitoring the network efficiency, factorized convolutions are capable of reducing the number of parameters in a network. Larger convolutions are replaced with smaller convolutions which leads to significant quicker training. Then, to reduce parameters, it factorizes to asymmetric convolution. During training, some loss is generated due to small convolution neural network injection between layers. This process is known as an

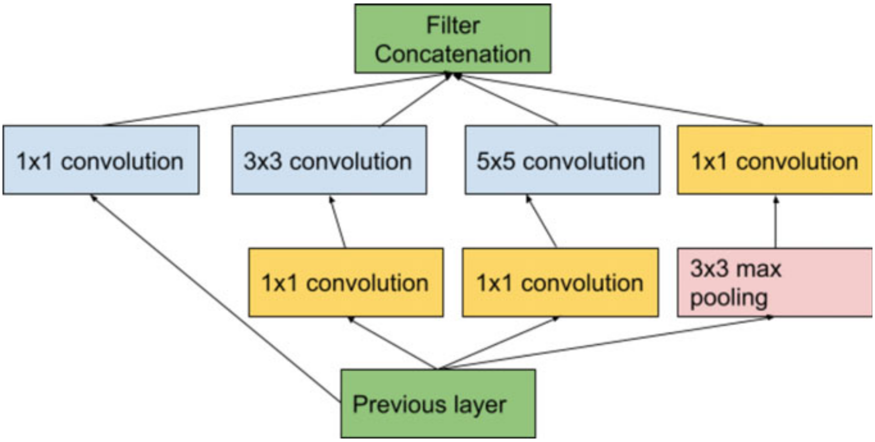


Fig. 1 Inception V3 demonstrates the Inception cell from Inception V3 model architecture. This model has 48 convolutional layers. Factorized convolutional layers, dimension reduction, and parallelization techniques are used in this method. We factorize the bigger parameter into smaller convolution and factorization into asymmetric convolution

auxiliary classifier which is used as a regularizer in Inception V3. Here, pooling techniques are commonly used to reduce grid size [18]. Figure 1 shows the Inception cell of Inception V3.

ResNet50 ResNet50 consists of 48 convolutional layers, MaxPool layer, and average pool layer. It introduced a technique known as residual mapping. Instead of expecting a few stacked layers to match a specified underlying mapping, the residual system allows these layers to intentionally fit a residual mapping. During training, the residual neural network has less filters and complexity than other models. Once a shortcut connection is added, the network is transformed into its residual counterpart and this shortcut connection is used to execute mapping, padding the dimensions with zeros. This does not add new parameters to the model to allow it to perform exceptionally well with very deep layers [19]. ResNet50 architecture is displayed in Fig. 2.

VGG16 VGG16 has 16 layers. For convolution, it always uses a 3×3 Kernel and the maximum pool size is 2×2 [20]. The image is translated using a series of convolutional layers with the smallest possible field size to capture notions such as left or right, center, and up or down including 1×1 convolution filters considered as linearly change input channels followed by nonlinearity. The spatial resolution of the layer input is retained after convolution. Five max pooling layers follow part of the convolution layers to perform spatial pooling. A stack of convolutional layers with varying depths in different architecture is followed by three fully connected layers. The softmax activation function is applied to the last fully connected layer. [21]. VGG16 is displayed in Fig. 3.

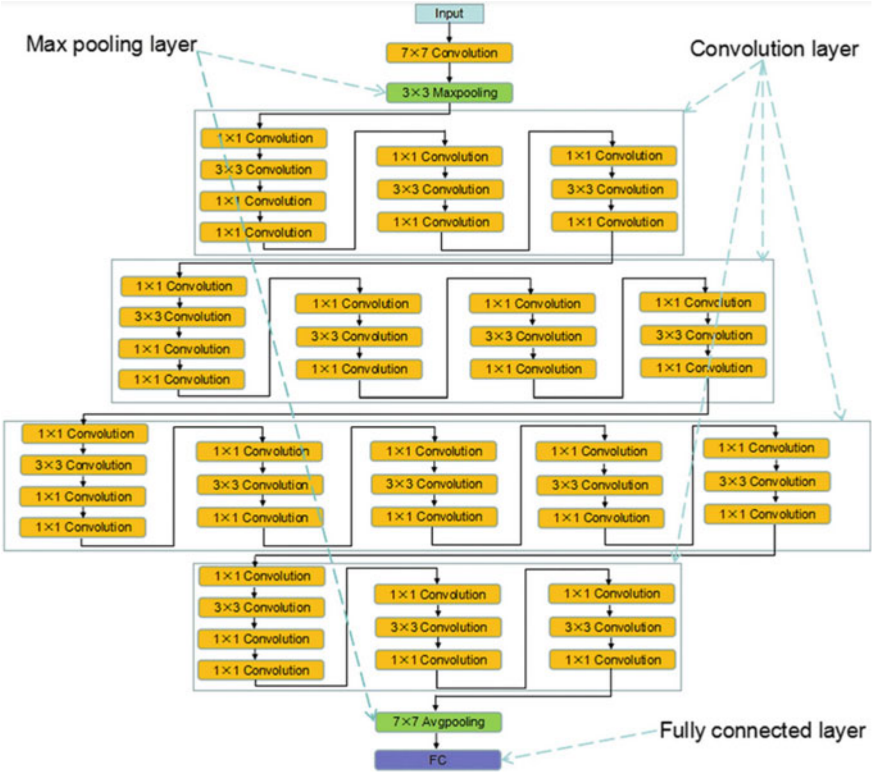


Fig. 2 ResNet50 model architecture consists of 48 convolution layers, one MaxPool layer, and one average pool layer. It is used to skip connections between layers and adds the output of previous layers to the outputs of the stacked layers

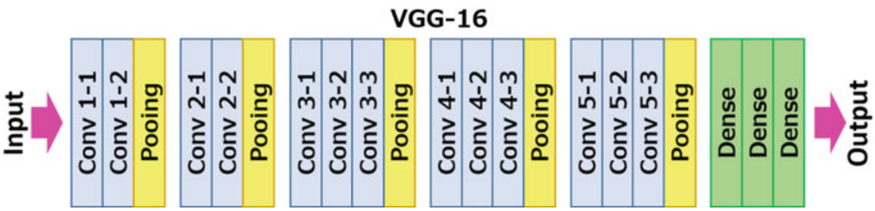


Fig. 3 VGG16 model architecture has 16 convolutional layers and an average 2D pooling layer. This model uses a set of weights that is pretrained on ImageNet

3.2 Fine-Tune Pretrained Model ResNet50 with Pruning

To prune the less important weight, we used magnitude-based weight prune in order to remove redundant weight, fine-tuning it to improve transfer learning performances.

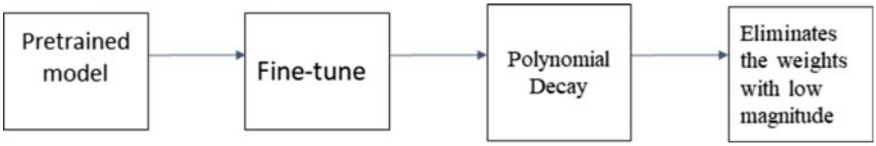


Fig. 4 Fine-tuning ResNet50 model with polynomial decay and eliminating the weights with low magnitude

We start pruning at the sparsity level at 50% and gradually increase it to train to reach 80% sparsity which means it starts from 50% and thus 80% weight is pruned. Each time it proceeds to reach final sparsity, it is mandatory to use step value. Thus, polynomial decay is used for reducing sparsity continuously and step value is required in order to reach the end sparsity. The proposed rice disease prediction model using the pretrained ResNet50 is illustrated in Fig. 4.

Generally, an arbitrary polynomial decay function could be defined as

$$g(n) = \sum_{j} \gamma_j n^j \tag{1}$$

Here, $g(n)$ denotes the monomial function for γ_js, j takes a series of values as $1, 2, 3 \ldots n$. But in case of polynomial decay, it can be expressed in Eq. 2

$$g'(n) = n^{\beta} \text{ for } \beta > 0 \tag{2}$$

where β denotes some exponent which are greater than zero and this condition demonstrates relative decay property (Fig. 5).

Algorithm 1 Algorithm 1 to implement Pruning and Fine-Tuning

Input: n,the number of pruning iterations , and X the dataset on which to train and finetune
$W \leftarrow initialize()$
$W \leftarrow trainToConvergence(f(X\ W))$
$M \leftarrow 1^{|w|}$
for i in 1 to n **do**
  $M \leftarrow prune(M, score(W))$
  $W \leftarrow finetune(f(X; MoW))$
end for
return M,W

We implemented Adam optimizer and sparse categorical cross-entropy in order to calculate categorical cross-entropy where the category index is represented by an integer encoder. Thus, it is capable of eliminating the weights with low magnitude and can be expressed as

$$J(w) = -y\log(\hat{Y}) \tag{3}$$

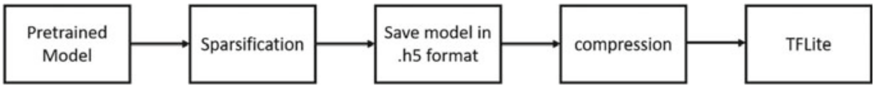


Fig. 5 Sparsified the VGG16 model after it is trained. We compressed the model and saved it in TFLite

where w denotes weights of the neural network and $\hat{Y}$ predicted level.

In this algorithm, neural network is implemented to train to convergence and for each parameter, network is pruned on the basis of scores and thus, usually fine-tuning is applied to recover accuracy of the whole network.

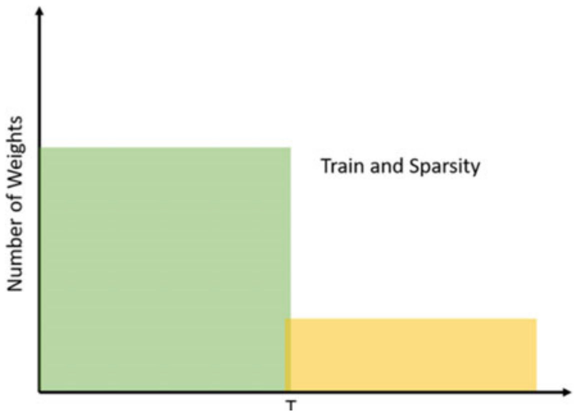
3.3 Pretrained Model VGG16 and Inception V3 with Pruning

In this part, we used a pretrained VGG16 model and aimed for sparsity that helps to decrease the size of the neural network. Pruning may lead us to set a maximum number of weights as zero and resulting the model having the same size as before. Therefore, compression algorithm is useful to be applied on it. It is important to note that restoring the previous model with sparse weight is possible after the model is compressed.

The train-then-sparsifying has sparsification of fully trained model that is denoted by yellow color in Fig. 6. It needs to proceed till convergence which is achieved using a number of iterations called T. Figure 6 illustrates standard dense training procedure and sparsification of mode.

Sometimes, training before sparsifying brings reduction in final accuracy. It aims at improving generalization and performance for inference and training, whereas sparsifying during training might lead the model to overfit.

Fig. 6 Train-then-sparsifying brings convergence by means if T iterations



3.4 Combination of Pruning and Quantization

In order to combine pruning and quantization, we take the pruned models and apply post-training quantization so that model size is reduced. The quantization process changes floating-point weights to integer. The activation is quantized dynamically during interference, and it is not possible to retrain these quantized weights. Figure 7 depicts the process of pruning followed by quantization. It is noteworthy that this pipeline has achieved a 10x model size reduction. Figure 7 depicts pruning followed by quantization. It is noteworthy that this pipeline has achieved a 10x model size reduction.

Fig. 7 Pruned model and therefore quantization applied to reduce the size of the parameters and result is converted into TF Lite

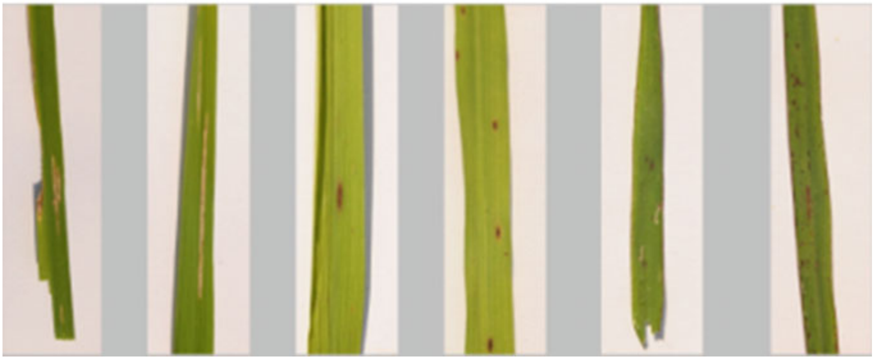


Fig. 8 Three types of rice diseases named bacterial leaf blight, brown spot, and leaf smut

4 Experiments

4.1 Dataset

The dataset we used is about rice plant diseases and was developed by Harshadkumar et al. [9] which is publicly accessible. It contains 120 images including 40 images of each disease and is classified into three different groups called leaf smut, brown spot, and bacterial leaf blight, respectively. These images consist of white background and captured in direct sunlight, and it is worth noting that all of the photographs are in jpeg format and the dimensions of all images are 2848 $\times$ 4288 pixels. We resize them to 224 * 224 pixels from 150 * 150 pixels. Figure 8 illustrates a few of the samples from the dataset.

Typically, these disease affect plant leaves, and the variations are mostly in color, texture, and shape. Here, symptoms might contain shaped-based variations such as elongated lesions on the leaf tip, round to oval shape, and tiny spots throughout the leaf, and the color might vary from yellow to white, reddish brown to dark brown or reddish brown.

4.2 Data Augmentation

Data augmentation is capable of boosting performance in classification tasks, and it could be made by patch rotation, mirroring, horizontal flip, vertical flip, zoom, and shear augmentation [22, 23]. We augment our dataset first and then train our model with augmented data. Table 1 illustrates data augmentation which is applied before train data in order to boost up the model performance.

Here, Fig. 9 depicts both images before and after augmentation.

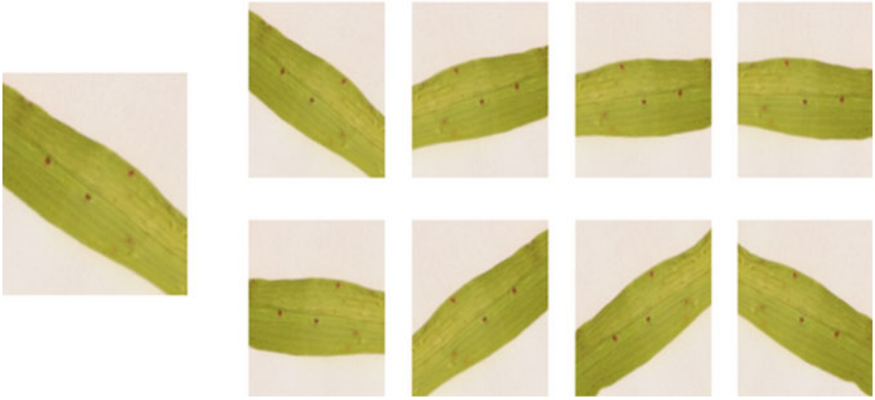


Fig. 9 Image before performing augmentation (**a**) image after augmentation is demonstrated in (**b**)

4.3 Baseline Method

We compare our proposed approach against a SVM classifier, and our model is trained on a dataset composed of 120 images with three categories. In this baseline paper [9], they performed detection and classification for rice disease prediction and applied four types of process to remove background and three types of segmentation such as Otsu's segmentation, LAB color space-based K-means clustering, and HSV color space-based K-means clustering. They created three classification models with 88, 72 and 44 features, respectively, and thus, they achieved 93.33% reported accuracy.

All our models were trained on a dataset when the training dataset is augmented. All models were trained and pruned. A version 00 Inception V3 was also quantized that led to models that are 3x–10x smaller than the original versions.

4.4 Training Procedure

To determine the most effective and efficient model, we performed data augmentation and then the image was resized to 224 × 224 images. We train it by choosing batch_size = 32 and ten epochs for every pretrained model. The Adam optimizer,

Table 1 Augmentation table

Augmentation properties	Parameters
shear_range	5.0–10.0
zoom_range	[1–0.2, 1+0.2]

which is an adaptive optimizer, is used with a sparse categorical cross-entropy loss function.

4.5 Implementation Details

All models were trained on Google Colab with Keras TensorFlow backend.

4.6 Evaluation Metrics

To develop an efficient model for rice disease prediction, we choose a series of pre-trained models. It can be assessed by evaluation metrics in comparison with baseline methods [24–26]. We use accuracy as a measure of correctness with respect to ground truth and floating-point operations (FLOPS) as a way of measurely efficiency.

Accuracy is a great measure for evaluating the correctness performance. It refers to correct prediction to all observations.

Therefore, FLOPS stands for floating-point operations per second, and it calculates the number of operations required to run any model [27]. It is noteworthy that FLOPs are used to freeze deep neural networks.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

5 Result and Discussion

In this paper, we explored efficient pretrained models such as VGG16, ResNet50, and Inception V3 which are developed further to detect rice disease prediction from rice leaves. Here, the best model performance of our model is 97% in terms of accuracy. Figure 10 demonstrates evaluation metric accuracy performance before and after model pruning.

In previous work, the reported accuracy of their model was 93.33%. It is evident that our models could perform inference within a very short time. We compared our result with the best past outcome. It is demonstrated that ResNet50 has performed tremendously better in comparison with the previously published results. Figure 11 depicts the result comparison among them.

Parameter reduction for all these models is depicted in Table 2.

Furthermore, we showed that we attempt to reduce the number of FLOPS for efficient and better performance. Inception V3 needs 8200 FLOPS whereas our developed model requires 7200 flops. Thus, it is worth mentioning that our proposed

Table 2 Parameters reduction for all models (VGG16, ResNet50, Inception V3, and Inception V3 with quantization)

Model	Parameters reduction
VGG16	29,953,064–4,978,883
ResNet50	47,568,043–262,659
Inception V3	45,655,069–22,818,979
Inception V3 with quantization	22,853,411–22,818,979

approach is much more efficient. Figure 12 shows the comparative performance in terms of FLOPS.

6 Conclusion

In this research study, we presented multiple pretrained models in order to detect three rice plant diseases including bacterial leaf blight, brown spot, and leaf smut and construct these models for efficient computation processes. We choose to augment our dataset followed by training with all pretrained models. We reveal through extensive experiments that our best performing model ResNet50 achieved 97% accuracy which outperform the 93.33% accuracy of of previously published methods. We conducted a series of experiments with model pruning including fine-tuning, pruning

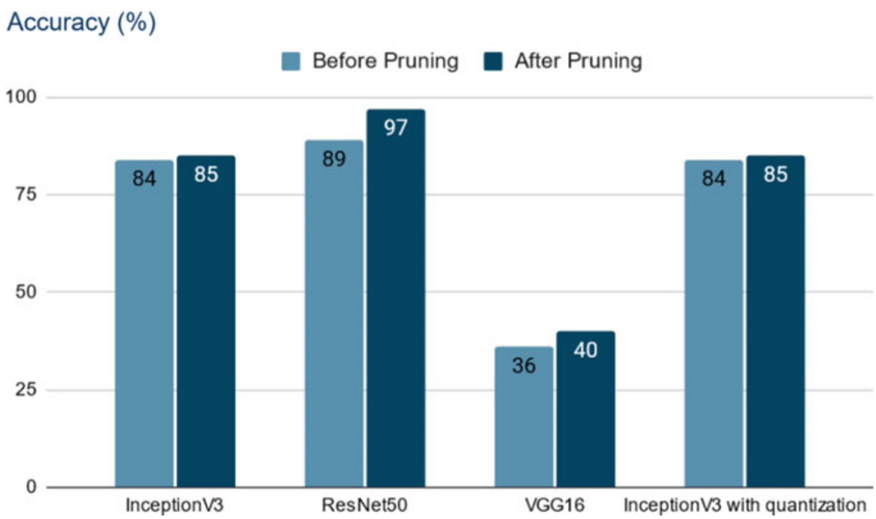


Fig. 10 Performance of ResNet50, VGG16, Inception V3, and Inception V3 with quantization using accuracy as evaluation metrics on our dataset

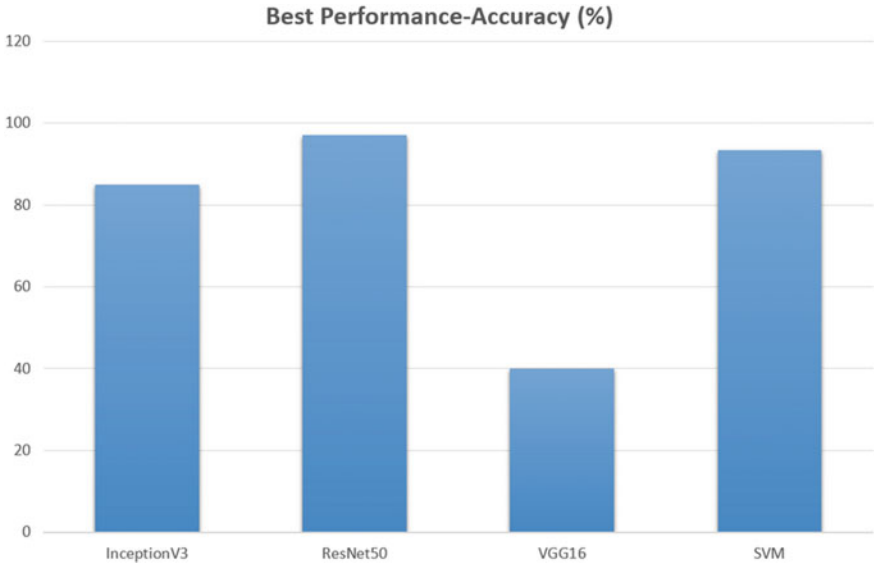


Fig. 11 Comparison with previous best results and ResNet50 outperforms the past outcome

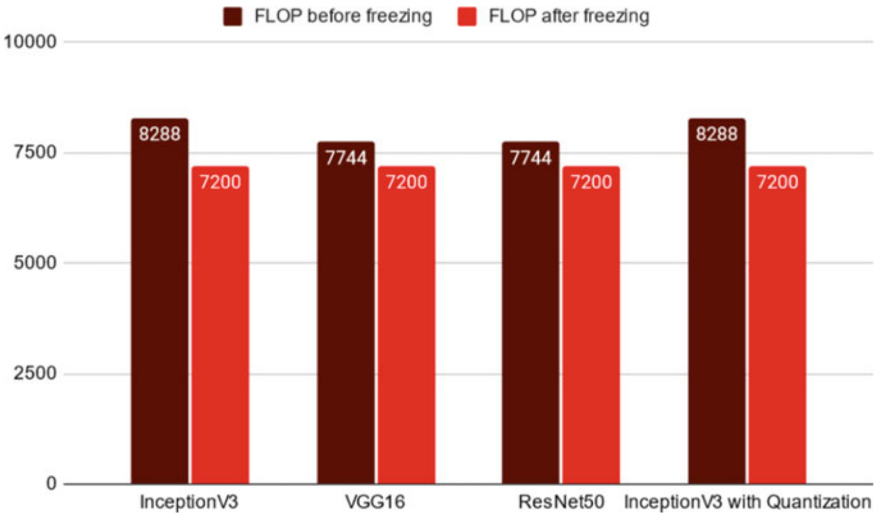


Fig. 12 Evaluation metric FLOPS and Inception V3 achieved the best result

with quantization and compression in order to reduce model size. Also, it is notable that our model Inception V3 requires only 7200 FLOPS that contribute to enhancing computational efficiency.

For future work directions, we would like to discover more models like EfficientNet and DenseNet with various types of datasets. Also, we could explore many other pruning techniques such as penalty-based pruning or recoverable pruning. We also plan to deploy it on mobile devices and build an autonomous system to detect rice disease.

References

1. Al Bashish, D., Braik, M., Bani-Ahmad, S.: A framework for detection and classification of plant leaf and stem diseases. In: 2010 International Conference on Signal and Image Processing, pp. 113–118. IEEE (2010)
2. Weizheng, S., Yachun, W., Zhanliang, C., Hongda, W.: Grading method of leaf spot disease based on image processing. In: 2008 International Conference on Computer Science and Software Engineering, vol. 6, pp. 491–494. IEEE (2008)
3. Gianessi, L.P.: Importance of pesticides for growing rice in south and south east Asia. Int. Pest. Benefit Case Study **108** (2014)
4. Mohan, K.J., Balasubramanian, M., Palanivel, S.: Detection and recognition of diseases from paddy plant leaf images. Int. J. Comput. Appl. **144**(12) (2016)
5. Suresha, M., Shreekanth, K., Thirumalesh, B.: Recognition of diseases in paddy leaves using knn classifier. In: 2017 2nd International Conference for Convergence in Technology (I2CT), pp. 663–666. IEEE (2017)
6. Suman, T., Dhruvakumar, T.: Classification of paddy leaf diseases using shape and color features. IJEEE **7**(01), 239–250 (2015)
7. War, H., Su, C.: Development of paddy diseased leaf classification system using modified color conversion. Int. J. Softw. Hardw. Res. Eng. **6**(8), 26–32 (2018)
8. War Htun, K., Htwe, C.S.: Development of paddy diseased leaf classification system using modified color conversion. Int. J. Softw. Hardware Res., Eng. (2018)
9. Bandara, D., Mayurathan, B.: Detection and classification of rice plant diseases using image processing techniques (2021)
10. Akila, M., Deepan, P.: Detection and classification of plant leaf diseases by using deep learning algorithm. Int. J. Eng. Res. Technol. (IJERT) **6**(07) (2018)
11. Rahman, C.R., Arko, P.S., Ali, M.E., Khan, M.A.I., Apon, S.H., Nowrin, F., Wasif, A.: Identification and recognition of rice diseases and pests using convolutional neural networks. Biosyst. Eng. **194**, 112–120 (2020)
12. Agus, F., Ihsan, M., Khairina, D.M., Candra, K.P.: Expert system for rice plant disease diagnosis (2019)
13. Yang, W., Chen, J., Chen, G., Wang, S., Fu, F.: The early diagnosis and fast detection of blast fungus, magnaporthe grisea, in rice plant by using its chitinase as biochemical marker and a rice cdna encoding mannose-binding lectin as recognition probe. Biosens. Bioelectron. **41**, 820–826 (2013)
14. Verma, T., Dubey, S.: Optimizing rice plant diseases recognition in image processing and decision tree based model. In: International Conference on Next Generation Computing Technologies, pp. 733–751. Springer (2017)
15. Xiao, M., Ma, Y., Feng, Z., Deng, Z., Hou, S., Shu, L., Lu, Z.: Rice blast recognition based on principal component analysis and neural network. Comput. Electron. Agricul. **154**, 482–490 (2018)

16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)
17. Inceptionv3 (Sep 2021). https://en.wikipedia.org/wiki/Inceptionv3
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. A guide to resnet, inception v3, and squeezenet (Sep 2021). https://blog.paperspace.com/popular-deep-learning-architectures-resnet-inceptionv3-squeezenet
20. Vgg 16 architecture, implementation and practical use (Sep 2021). https://medium.com/pythoneers/vgg-16-architecture-implementation-and-practical-use-e0fef1d14557
21. Vgg 16 (Sep 2021). https://neurohive.o/en/popular-networks/vgg16
22. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A.: Classification of breast cancer histology images using convolutional neural networks. PloS One **12**(6), e0177544 (2017)
23. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 289–293. IEEE (2018)
24. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)
25. Perez, F., Vasconcelos, C., Avila, S., Valle, E.: Data augmentation for skin lesion analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 303–311. Springer (2018)
26. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imag. **36**(4), 994–1004 (2016)
27. Blalock, D., Ortiz, J.J.G., Frankle, J., Guttag, J.: What is the state of neural network pruning? arXiv preprint arXiv:2003.03033 (2020)

Global Warming and Bangladesh: A Machine Learning Approach to Analyze the Warming Rate Utilizing Neural Network

Tamal Joyti Roy and Md. Ashiq Mahmood

Abstract Bangladesh is among the few nations on this planet which will confront the lethal results of a worldwide temperature alteration. Global warming is the crisp product of climate change and it is real. To predict the global warming rate of Bangladesh, we have used eight machine learning algorithms with a dataset consisting of about five thousand data. We have collected data from the University of Dayton—Conversational Guard Organization from the Average Daily Temperature Library. We have proposed neural network systems for calculating the nonexistent data of the temperature dataset. We have also performed a correlation test for both test and train datasets and that outcome was significant. Our data period is from 1995 to 2010. The accuracy of our machine learning experiment shows that the Naive Bayes algorithm can predict the global warming consequences in Bangladesh with over 99 percent accuracy. Global warming is happening, sea level of the Bay of Bengal is rising, our experiment has showed that global warming is affecting Bangladesh, and the outcome of that will be immense. It is the right time we should start our awareness of climate change.

Keywords Naive Bayes · Simple logistic · Neural network · Machine learning · Climate change · Computer vision · Global warming

1 Introduction

The slogan "global warming" has become aware to many people as one of the vital environmental issues of our day. The decades of 80 and 90 were strangely warm [1]. Concerning global ordinary near-surface air temperature, the year 1998 was the hottest in the contributory record, and the nine hottest years in that record have occurred since the 90s. Global warming is real, although there has been a debate that many people think of global warming as a hoax. Long before Homo sapiens arrived on earth, there were animals and birds. Earth had seen five massive extinctions [2].

T. J. Roy (✉) · Md. Ashiq Mahmood
Khulna University of Engineering and Technology, Khulna, Bangladesh
e-mail: tjroy13june@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_2

Storms or tornado are by no means the only weather and climate excesses that cause catastrophes [3]. We have collected 15 years of temperature data from Bangladesh [4]. There is sufficient proof for huge deviations in ocean smooth in the time of the earth's past, in the time of the warm period before the start of the last Ice Age; about millions of years ago, the worldwide regular temperature was a little warmer. Linking machine learning styles for the prediction of spatially unambiguous live around was compared the predictive accuracy and efficiencies of five separate supervised machine learning algorithms, testing various sample sizes and quantity findings, and control the most fruitful and precise architecture for approximating spatially outer life cycle GW and EU influences at the minimum scale, with corn production in the US's Midwest region as a case study. This study likened the prognostic exactitudes, efficiencies of five distinct supervised machine learning algorithms, various testing sample sizes, and feature selections. Results showed that the gradient-increasing reversion tree architecture built in an approximation of 4000 records of monthly forecasting features produced the maximum predictive result with cross-authentication standards of 0.8 for the reproduction cycle GW influences. Most of the studies did not count the South Asian countries temperature rising, and we have tried to work in this area. Never before Bangladesh global warming rate predictions analysis had not been done. Some studies used the phonological models tied with ML methods for the detection of temperature changes in the atmosphere. One study has analyzed global warming with the help of machine learning which comprises linear regression, lasso, SVM, and random forest to shape the form of the model to authenticate the global warming process of the earth and categorizing issues contributing the global warming [5]. Our study mainly has focused on establishing a machine learning prediction on global warming from Bangladesh's perspective (Fig. 1).

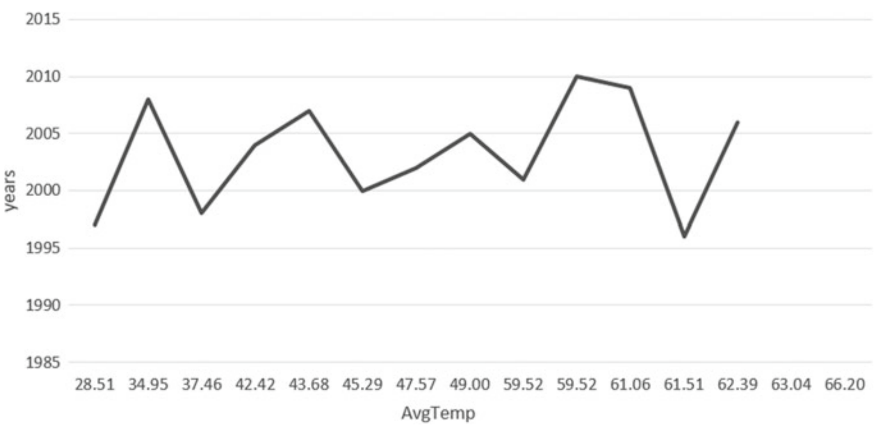


Fig. 1 Bangladesh average temperature between 1995 and 2010

2 Related Works

For a sustainable energy future, there is another study [6] that focused on the renewable energy resources bonded with machine learning technologies theoretically. Prediction of spatial and temporal variability in vegetation produces used linear regression and machine learning models, and machine learning showed best probabilistic results [7]. Metaheuristic techniques are applied for predicting photostatic power generation with the help of machine learning classifiers [8]. A survey had been made, and the outcomes were pretty good [9]. For building climate change sensibility simulations, another machine learning approach had been conducted with ozone parameters [10]. They showed key development procedure to additional improvement and range of the possibility of ozone parameterizations that based on ML. CO_2 emissions prediction was conducted with this study [11] where clustering and machine learning combination was used and there were adaptive neuro-fuzzy interface systems. Machine learning techniques are used for identifying the storage behavior of ocean gyre [12], as light-driven decreases in chlorophyll can be related to fixed or even above photosynthesis. A study was conducted to model the climate, climate change, and extreme events [13]. Where accuracy varies from 0 to 20% [14], understandings into proceeding process-founded models for prominent N dynamics and greenhouse gas releases are using a machine learning approach. This study [15] showed a relation of drought in China using machine learning classifiers. A monthly prediction of weather forecast was used by some machine learning classifiers like support vector regression and multi-layer perceptron [16]. For predicting the stream water temperature prediction, six machine learning algorithms and classifiers were used including linear regression, random forest, extreme gradient boosting forwarding neural networks, and two types of repeated neural networks [17].

3 Proposed Methodology

3.1 Data Preprocess

First of all, we have prepossessed the datasets. There were 2356 missing data, denoted with "−99"; we have converted those with assigning value 1. After that [18], procedure was followed. All the data were in Fahrenheit ($33.8 Fahrenheit = 1 degree Celsius$). Total data instances were 5810, from there 2356 were missing quantities. We have chosen the past dataset because the availability of new data was rare and most of the datasets did not contain enough preprocessed data. We needed to just predict the global warming impact scenario; that is why we have made both a neural model and performed machine learning approaches.

Table 1 Sample dataset

Year	Avg	Warming (yes or no)	Year	Avg	Warming (yes or no)
1995	64.8	0	2006	79	0
1995	58.4	0	2006	79.6	0
1995	0	0	2006	80	0
1995	59.6	0	2006	79.9	0
1995	64.9	0	2006	79.9	0
1995	0	0	2006	80.4	1
1995	70.9	0	2006	80.7	1
1995	70.3	0	2006	82	1
1995	64.3	0	2006	82	1
1995	66.6	0	2006	0	0
1995	70.6	0	2006	0	0
1995	65.9	0	2006	86	1
1995	67	0	2006	80.6	1
1995	60.4	0	2006	0	0
1995	68	0	2006	75.6	0
1995	66.7	0	2006	78	0
1995	0	0	2006	79	0
1995	65.5	0	2006	0	0
1995	66.2	0	2006	80.9	1
1995	0	0	2006	83.6	1
1995	56.7	0	2006	78	0
1995	66	0	2006	79.4	0
1995	62.4	0	2006	82	1
1995	0	0	2006	0	0
1995	0	0	2006	79.8	0

3.2 Correlation

We have done a correlation test with the average temperature and the global warming dataset (we have assigned the data which are more than 80 °F to the class "Yes" and less than that to "No"). Table 1 shows the two-tailed correlations between the avg and warming. There have been a significant relations at 0.660, the value of N = 5810. Without the value of N remains 3454. We have split our data with an 80:20 ratio, which means for training, we used 80 percent data, and for the test, we used 20 percent data. Figure 2 shows the flowchart diagram of our experimental analysis.

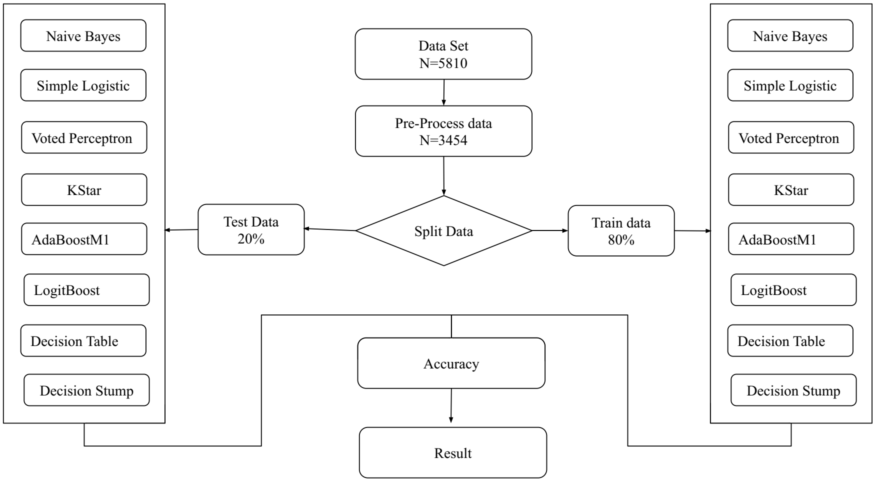


Fig. 2 Flow diagram of experimental analysis

3.3 Calculation

Neural Network: A neural network is a blends of algorithms that deeds to understand the underlying relations in a set of data by processing that satirists, the way the human mind functions. So neural networks denote to systems of neurons, both organic and artificial in nature. The data were calculated through a 24 h strategy. For example, getting the average temperature Eq. 1 is followed, which is a simple equation. We have used eight machine learning algorithms for predicting our results. Algorithms are Naive Bayes, simple logistic, voted perceptron, KStar, AdaBoostM1, LogitBoost, decision table, decision stump. We have split our data with an 80:20 ratio, which means for training, we used 80% data, and for the test, we used 20 percent data. Figure 2 shows the flowchart diagram of our experimental analysis.

We have made a new neural network system that produces some accuracy with the help of deep learning without missing any data as we had some missing data in the first place, so we came up with this idea. Our dataset is nonlinear which means the general algebra theory would not work.

$$\text{AvgTemp} = \frac{\text{mintemp} + \text{maxtemp}}{2} \tag{1}$$

The multi-layer neural network system helped in this problem. We have supervised it. We added three layers in the input layer of our neural network model, temp1 dataset avg of the month, day, and week. Temp2 dataset avg of the month, day, and week, and lastly temp of n dataset avg of month, day, and week. We fed the neural network layers and output it with accuracy1, accuracy 2, and last of all accuracy n, and we have gotten the accuracy tested an average of temperature, month, day,

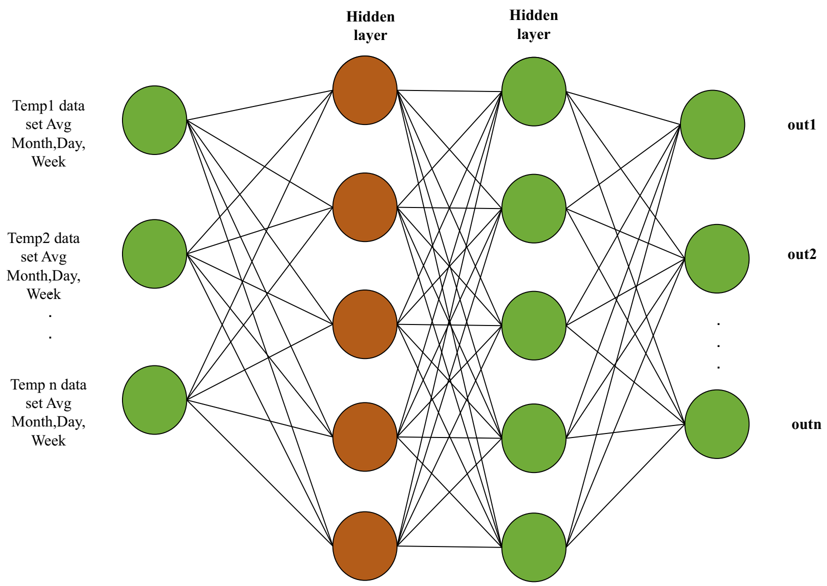


Fig. 3 Neural network system for accuracy testing

Table 2 Correlation

		Avg	Warming
Avg	Pearson correlation	1	0.66
	S (2)		0.000
	N	5810	5810
Warming	Pearson correlation	0.660	1
	S (2)	0.000	
	N	5810	5810

and week. And this helped to merge with the original dataset that was supposed to be missing. Figure 3 shows our neural network model. Table 2 shows the sample dataset. The year in that table represents every year in int number, month represents the numeric of the month; for example, January is assigned as 1, February assigned as 2, and so on. The mathematical notation of our neural network is:

$$\sum = (\text{AvgT} \times mt \times D) + (\text{AvgT}n \times mt \times D) \tag{2}$$

The gradient of the predicted value in the output layer denoted ask; now from Eq. 3, we are getting the results of the output layer:

$$\frac{\mathrm{d}(\mathrm{temp}n)}{\mathrm{d}k} = \frac{\mathrm{d}}{\mathrm{d}k}\beta(k) \tag{3}$$

$$= \frac{\mathrm{d}}{\mathrm{d}k}\left(\frac{1}{1+e^{-k}}\right)^2$$

$$= \beta\,(k) + (1 - \beta(k))$$

Getting data polished from the output layer, we again preprocessed it and started to test with the machine learning algorithms. Eight machine learning algorithms produced some kind of results. Our experiment was conducted with Bayesian, functions, meta, rules, and tree classifiers. Our accuracy was between 72% and 99% in both test and training datasets.

3.4 Algorithms

Naive Bayes: Based on Bayes proposition. It is not a sole algorithm but a clan of algorithms where all of them bit a common standard and every duo featuring is being classified as independent of each other.

Simple Logistic: It catches the notation that best forecasts the value of the Y variable for each value of the X variable. This makes logistic regression different from linear regression.

Voted Perceptron: The voted perceptron technique is founded on the perceptron algorithm of Rosenblatt and Frank. The algorithm takes lead data that are linear and divisible with large margins.

KStar: It is a situation-driven classifier, that is, the dataset; a test case depends upon the class of those training sets similar to it, as resolute by some same function.

AdaBoostM1: Which start with suitable for a classifier on the real dataset and then convulsions others duplicates of the classifier on the same dataset but where the weights value of erroneously classified cases are managed such that succeeding classifiers focus more on challenging cases.

LogitBoost: LogitBoost and AdaBoost are nearly in the sense that both do stabilizer logistic regression. The difference is that AdaBoost focuses on exponential loss.

Decision Table: It is used to stipulate what movements should be done under minimal circumstances, which are generally articulated as true (T) or false (F).

Decision Stump: It contains of a one-near decision tree; it is a decision tree with one inner lump that is directly connected to station nodes. A decision stump has varieties of the predictions based on the marks of just a sole input.

4 Experimental Results

Our datasets were divided into two subsets, 80% training data and 20% test data. We did not perform any cross-validation. We applied the Naïve Bayes first, then simple logistic, and followed by voted perceptron, KStar, AdaBoostM1, LogitBoost, decision table, and decision stump classifier algorithms for predicting the outcomes. The Naïve Bayes gave more accuracy in predicting with 99.93% accuracy in training data and 98.06% in training data. In insights, connection or reliance is any factual relationship, if causal, between two irregular factors or bivariate information. In a sense, connection is any measurable affiliation; however, it usually alludes to how much a couple of factors that are straightly related can be denoted for correlation matrix. The correlation matrix showed the linear relationship among the datasets. The simple logistic gave 99.75 in the training set and 99.59 in the test dataset. Voted perceptron 99.91 % in training and 94.92 in the test. Other algorithms KStar, AdaBoosM1, LogitBoost, decision table, and decision stump gave 77.34, 72.24, 73.56, 99.91, and 72.23 percent in training dataset and 75.68, 72.43, 73.70, 97.91, and 72.44 in test dataset. Figure 4 shows the training and test datasets accuracy. Figure 5 shows the true, false positive rate, and F-measure of our algorithms. Figure 6 shows the visualization of neural network deep learning accuracy. The data samples are used to deliver a clear outcome of a model or architecture or framework and are suitable for the training dataset in time of alteration model hyperparameters.

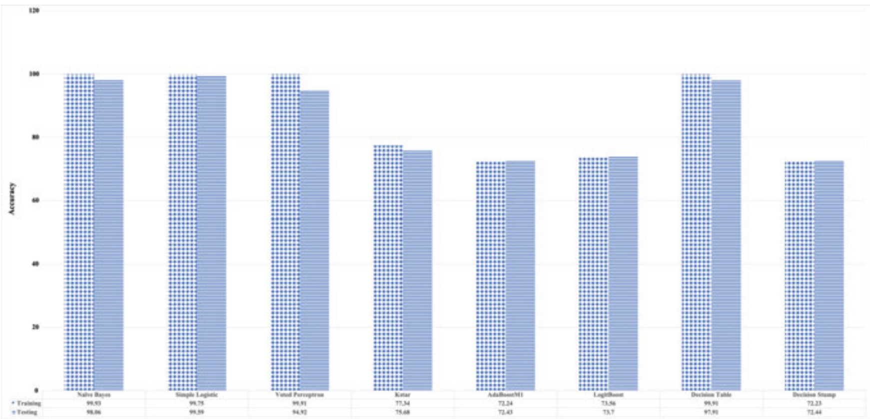


Fig. 4 Accuracy of training and test datasets

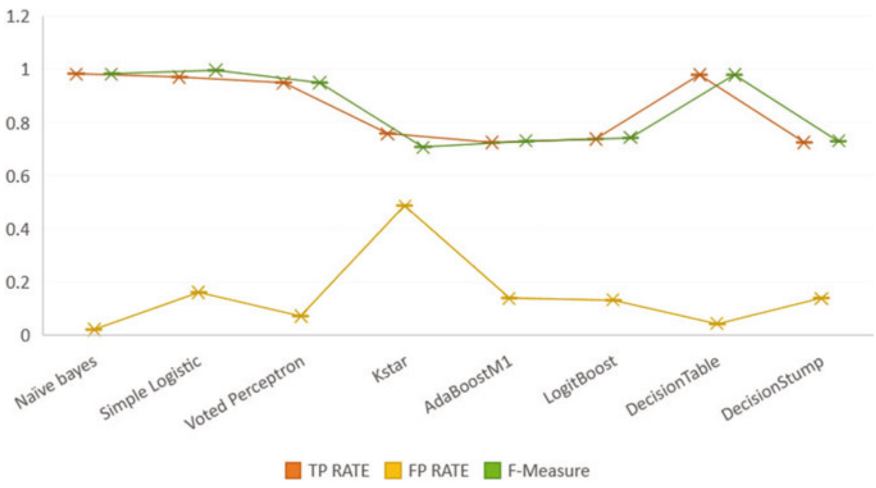


Fig. 5 Accuracy class measurement TP, FP, F-measure

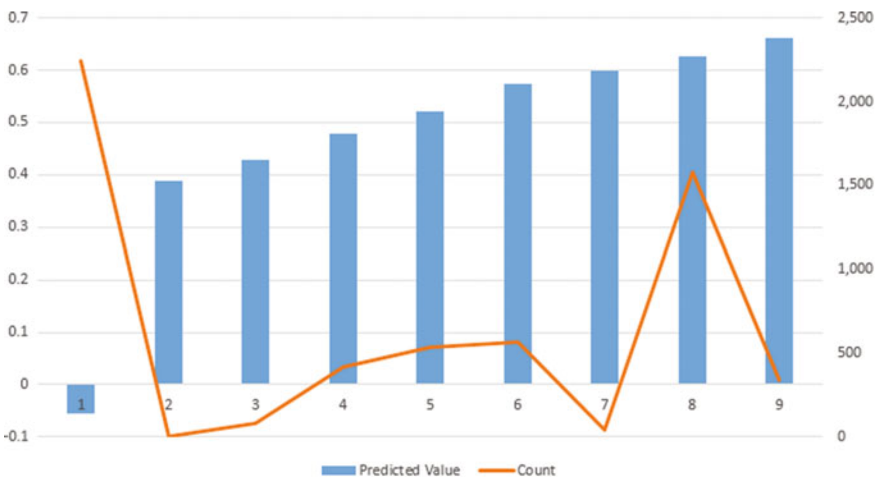


Fig. 6 Neural network deep learning accuracy visualization

When any assessment develops influenced by rating on the confirmation dataset is incorporated into the model configuration (Table 3).

In this case, the avg temperature, warming, and year showed a linear relation with the matrix (Table 4).

A false positive is a consequence when any architecture or model or method incorrectly calculates the positive class. And it becomes false negative and is a consequence when any architecture or model or method incorrectly predicts the class.

A true positive is a consequence when any architecture or model or method appropriately calculates the positive class. Likely, a true negative is a consequence when

Table 3 Receiver operating characteristic and precision–recall

Machine learning algorithms	ROC	PRC
Naïve Bayes	0.9	0.9
Simple logistic	0.98	0.98
Voted perceptron	0.94	0.95
KStar	0.96	0.93
AdaBoostM1	0.79	0.86
LogitBoost	0.835	0.884
Decision table	0.96	0.95
Decision stump	0.79	0.86

Table 4 Prediction table

Warming	Predicted value	Count
0	−0.0562	2243
0	0.3893	6
0	0.4285	83
0	0.4793	417
0	0.5222	538
0	0.5746	569
1	0.5995	42
1	0.6263	1576
1	0.661	336

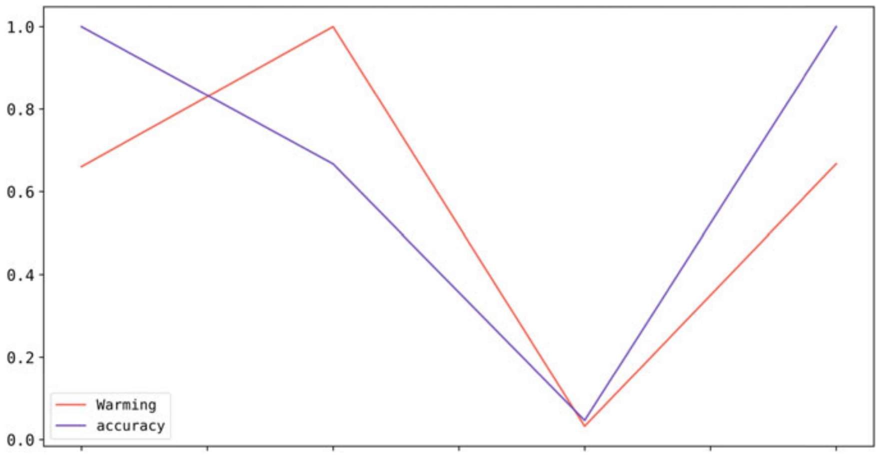


Fig. 7 Predicting results obtained from ML algorithms

any architecture or model guesses the class. The correlation matrix showed the linear relationship among the datasets. ROC area is distinct from the FPR and TPR. Two axes coordinates correspondingly portray comparative between true positive and false positive. The TPR is correspondent to understanding, and FPR is always equal to 1. F-measure delivers a way to merge two things—precision and recall into a single quantity that detentions two possessions, of course, precision or recall tell the full section. Having good precision values with fewer recall values, or interchangeably, less precision with good recall, it provides a method to prompt with concerns and single score. The neural network system followed Eq. 3 for calculating the accuracy. The accuracy from neural network shown in Fig. 7 is that accuracy obtained after 5899 iterations, all the accuracies are compared with the machine learning algorithms, and the results vary from 60 to 99%. Hence, our both models predicted almost the same results in terms of accuracy, but the machine learning model needed less time and space than the deep learning of neural network system.

5 Conclusion

There are ample evidences for large variations in ocean uniformity in earth's history, during the warm period before the start of the last Ice Age, approximately millions of years ago, when the global average temperature was a little warmer. The accuracy of machine learning algorithms has showed well-furbished outcomes, and neural network multi-layer system has been able to fix the missing values and output accuracy with more than five thousand iterations. The outcome of our experiment is that the Naïve Bayes, voted perceptron, and decision table's accuracy is almost more than 99%, but the true positive rate, false positive rate, and the F-measure showed that Naïve Bayes showed the best result among eight algorithms. In the future studies, our experiment can help to predict the global warming of other countries such as India, Sri Lanka, Pakistan, Afghanistan. One of the limitations of our study is that—other machine learning algorithms can be used to analyze the accuracy, multiple regression analysis can be placed for logistic results. Our model also can be used in ozone layer biased, where the ozone layer gases are not detected properly. The bright side is if we start the conscious process about global warming, the whole warming process will slow down.

References

1. The Public and Climate, cont. https://history.aip.org/climate/public2.htm. Accessed 25 Jun 2021
2. Benton, M.J.: Hyperthermal-driven mass extinctions: killing models during the Permian-Triassic mass extinction. Philos. Trans. Royal Soc. A: Math. Phys. Eng. Sci. **376**(2130) (Royal Society Publishing) (2018). https://doi.org/10.1098/rsta.2017.0076

3. Arnell, N.W., Lowe, J.A., Challinor, A.J., Osborn, T.J.: Global and regional impacts of climate change at different levels of global temperature increase. Clim. Change **155**(3), 377–391 (2019). https://doi.org/10.1007/s10584-019-02464-z. Aug
4. Climate Change & its Impacts on Bangladesh | NCDO. http://www.ncdo.nl/artikel/climate-change-its-impacts-bangladesh. Accessed 25 Jun 2021
5. Zheng, H.: Analysis of global warming using machine learning. Comput. Water, Energy, Environ. Eng. **07**(03), 127–141 (2018). https://doi.org/10.4236/cweee.2018.73009
6. Gu, G.H., Noh, J., Kim, I., Jung, Y.: Machine learning for renewable energy materials. J. Mater. Chem. A **7**(29), 17096–17117 (Royal Society of Chemistry) (2019). https://doi.org/10.1039/c9ta02356a
7. Leng, G., Hall, J.W.: Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. Environ. Res. Lett. **15**(4), 44027 (2020). https://doi.org/10.1088/1748-9326/ab7b24. Apr.
8. Akhter, M.N., Mekhilef, S., Mokhlis, H., Shah, N.M.: Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. IET Renew. Power Generation **13**(7), 1009–1023 (Institution of Engineering and Technology) (2019). https://doi.org/10.1049/iet-rpg.2018.5649
9. Lai, J.P., Chang, Y.M., Chen, C.H., Pai, P.F.: A survey of machine learning models in renewable energy predictions. Appl. Sci. (Switzerland) **10**(17), 5975 (2020). https://doi.org/10.3390/app10175975
10. Nowack, P., Braesicke, P., Haigh, J., Abraham, N.L., Pyle, J., Voulgarakis, A.: Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. Environ. Res. Lett. **13**(10), 104016 (2018). https://doi.org/10.1088/1748-9326/aae2be. Oct
11. Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M., Cavallaro, F.: A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. J. Clean. Prod. **275**, 122942 (2020). https://doi.org/10.1016/j.jclepro.2020.122942. Dec
12. D'Alelio, D., et al.: Machine learning identifies a strong association between warming and reduced primary productivity in an oligotrophic ocean gyre. Sci. Rep. **10**(1), 1–12 (2020). https://doi.org/10.1038/s41598-020-59989-y. Dec.
13. O'Gorman, P.A., Dwyer, J.G.: Using machine learning to parameterize moist convection: potential for modeling of climate, climate change, and extreme events. J. Adv. Model. Earth Syst. **10**(10), 2548–2563 (2018). https://doi.org/10.1029/2018MS001351. Oct.
14. Saha, D., Basso, B., Robertson, G.P.: Machine learning improves predictions of agricultural nitrous oxide (N2O) emissions from intensively managed cropping systems. Environ. Res. Lett. **16**(2), 024004 (2021). https://doi.org/10.1088/1748-9326/abd2f3. Feb.
15. Zhang, R., Chen, Z.Y., Xu, L.J., Ou, C.Q.: Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China. Sci. Total Environ. **665**, 338–346 (2019). https://doi.org/10.1016/j.scitotenv.2019.01.431. May
16. Salcedo-Sanz, S., Deo, R.C., Carro-Calvo, L., Saavedra-Moreno, B.: Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. Theor. Appl. Climatol. **125**(1–2), 13–25 (2016). https://doi.org/10.1007/s00704-015-1480-4. Jul.
17. Feigl, M., Lebiedzinski, K., Herrnegger, M., Schulz, K.: Machine-learning methods for stream water temperature prediction. Hydrol. Earth Syst. Sci. **25**(5), 2951–2977 (2021). https://doi.org/10.5194/hess-25-2951-2021. May
18. Daily Temperature of Major Cities|Kaggle. https://www.kaggle.com/sudalairajkumar/daily-temperature-of-major-cities/discussion/160690. Accessed 23 Jun 2021

An EEG-based Intelligent Neuromarketing System for Predicting Consumers' Choice

Fazla Rabbi Mashrur, Khandoker Mahmudur Rahman, Mohammad Tohidul Islam Miya, Ravi Vaidyanathan, Syed Ferhat Anwar, Farhana Sarker, and Khondaker A. Mamun

Abstract Marketers use different marketing strategies to elicit the desired response from the target customers. To measure customer engagement, they usually conduct one-on-one interviews, surveys, broad polls, and focus group discussions. However, these are costly and sometimes unreliable. On the other hand, neuromarketing measures customer response to marketing stimuli by measuring the electrical activity of the brain and has the potential to address these drawbacks. which can be overcome by neuromarketing. In this work, we proposed a prediction algorithm that can identify consumer affective attitude (AA) and purchase intention (PI) from EEG signals from the brain. At first, the raw EEG signals are initially preprocessed to remove noise. After that, three features are extracted: time, frequency, and time frequency domain features. Wavelet packet transform is used to separate the EEG bands in time-domain feature extraction. Then, for feature selection, wrapper-based SVM-RFE is utilized. Finally, we use SVM to distinguish the classes in AA and PI. Results show that for SVM with radial basis function kernel performs better to classify positive and negative AA with an accuracy of 90 ± 4.33 and PI with an accuracy of 75 ± 2.5. So, EEG-based neuromarketing solutions can assist companies and organizations in accurately pre-

Supported by ICT innovation Fund, ICT Division, MoPTIT, GOB and Institute of Advanced Research (IAR), United International University (UIU).

F. R. Mashrur · K. A. Mamun (✉)
Advanced Intelligent Multidisciplinary Systems Lab, Institute of Advanced Research, United International University, Dhaka, Bangladesh
e-mail: mamun@cse.uiu.ac.bd

K. M. Rahman · M. T. I. Miya
School of Business and Economics, United International University, Dhaka, Bangladesh

R. Vaidyanathan
Department of Mechanical Engineering, Imperial College London, London, UK

S. F. Anwar
Institute of Business Administration, University of Dhaka, Dhaka, Bangladesh

F. Sarker
Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_3

dicting future consumer preferences. As a result, neuromarketing based-solutions have the potential to increase sales by overcoming the constraints of traditional marketing.

Keywords Consumer neuroscience · Neuromarketing · Electroencephalography · Consumer behavior · Pattern recognition

1 Introduction

Neuromarketing is the evaluation of neural or physical activity to gain insight into customers' preferences and intentions, which can help the marketers in product development, pricing, and advertising. The major goal of marketers is to display their product in such a way that it generates the desired response to their target customers. Companies use different marketing strategies such as product promotion, celebrity endorsement, and other offers to accomplish this. They usually conduct one-on-one interviews, surveys, broad polls, and focus group discussions to gauge consumer preferences or decisions in order to assess their success [1]. Nonetheless, even though these strategies are simple, they can be costly to implement. Furthermore, they produce results that may contain biases, making them appear unreliable [1]. Hence, utilizing the power of technology, autonomous customer choice prediction is required. As a result, neuromarketing, the merging of neuroscience, technology, and traditional marketing, is introduced.

Neuromarketing measures customer response to marketing stimuli by measuring the electrical activity of the brain, imaging, or other activity monitoring methods. According to prior studies, the frontal is primarily responsible for product decision-making and likeability [1]. There are several technologies that are used to observe brain activity during neuromarketing studies. Among all the technologies to discover brain activity, the electroencephalogram (EEG) is the most cost-effective, portable, and has the highest temporal resolution. As a result, the use of EEG in neuromarketing research has grown lately. Furthermore, to establish such a system in a real-life setting EEG is the most suitable option. Hence, in this work, we use EEG for the aforementioned reasons.

Several EEG-based neuromarketing tests are conducted to see how advertisement design affects customer decision-making and shopping behavior. The spatiotemporal connections of continuous EEG data sources [4] aim to model and simulate the patterns of brain functions produced by marketing stimuli. The findings indicate that negative framing gives more exposure, while positive framing created fewer difficulties in decision making and higher buying intentions, a key marketing problem. Machine learning (ML) is also used to accomplish automatic preference prediction identification. Yadava et al. [12] introduced a predictive ML model for the classification of "likes" and "dislikes". Jin et al. [5] examined the effects of online shopping, using ERP analysis, of a product description on the buying choice. Chew et al. [3] investigated virtual 3D shape preference from EEG signals using a machine learning classifier. While recent researches have demonstrated the benefits of EEG for neuromarketing applications, some open-ended issues need to be answered with more research. The efficiency potential of EEG signals to determine and forecast choice, the influence of stimulus material on desires, and buying decision-making processes need to be investigated further. To the best of the authors' knowledge, there is no study that explored ML framework from EEG signals using marketing stimuli such as endorsement and promotion. Moreover, previously in literature researchers have not studied consumer's buying decisions using ML. To investigate such a gap, here, we introduce a prediction algorithm that can identify consumer preference and buying decisions from EEG signals while marketing stimuli are being viewed. As it appears from extant literature, this is a leading and early study that predicts affective attitude (AA) for marketing stimuli and forecasts purchase intention (PI) of customer using ML algorithm. We also conduct comprehensive experiments on tuning the hyperparameter of the classification model.

Rest of the manuscript is organized as follows: Sect. 2 provides a description of the dataset as well as the theoretical background, Sect. 3 analyzes the obtained results and discusses them, and lastly, Sect. 4 concludes the paper by drawing conclusions about the results and possible future work related to this study.

2 Materials and Methods

Figure 1 demonstrates the block diagram of our proposed framework. In this section, we discuss about the participants, stimuli, and methodological description of this study (Fig. 2).

2.1 Participants

Five healthy volunteers with no history of neurological illness take part in this study. All participants provided informed consent prior to enrollment, in accordance with the Helsinki Declaration and the Neuromarketing Science and Business Association Code of Ethics.

2.2 Stimuli Description

In this study, we employ five different products in our work, each with its endorsement and promotion advertisement. A product endorsement is a type of advertising that has a favorable impact on buyers. In most cases, celebrities advocate a product in a real-life setting. However, in our work, we purposefully employ neutral endorsement to prevent biasing the participants. In promotion, we have given buy one get one offer

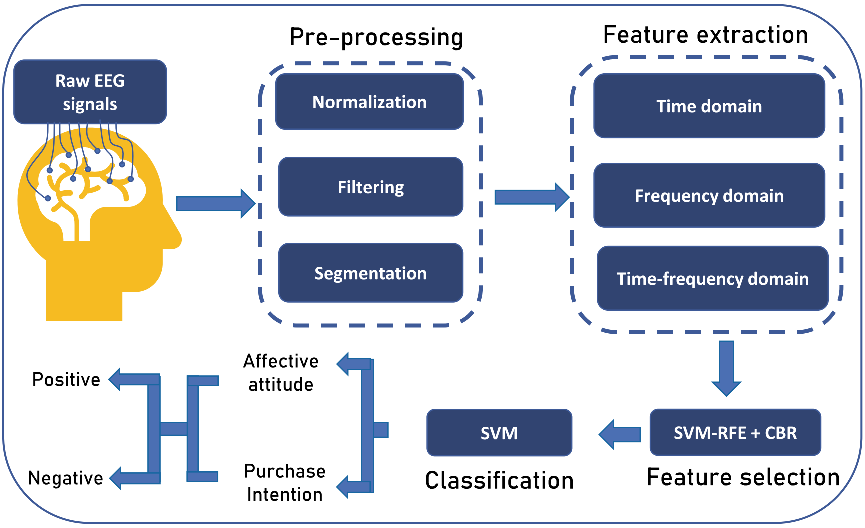


Fig. 1 The workflow of our proposed pipeline is depicted. The raw EEG signals are initially preprocessed to remove noise. After that, three feature are extracted: time, frequency, and time–frequency domain features. Then, for feature selection, wrapper-based SVM-RFE along with CBR is utilized. Finally, we use SVM to distinguish the classes in AA and PI

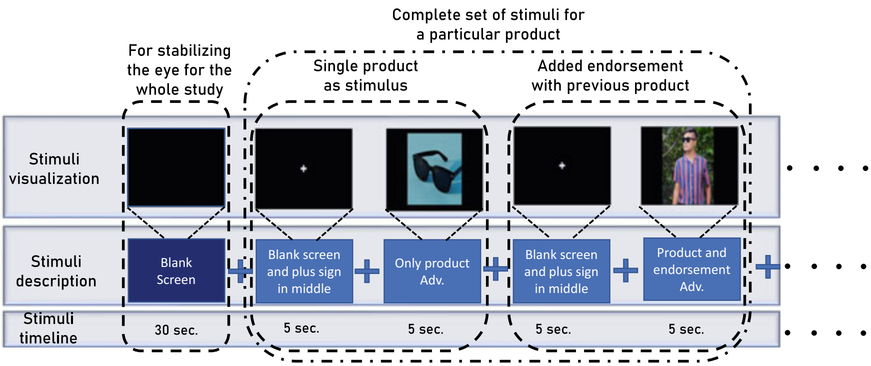
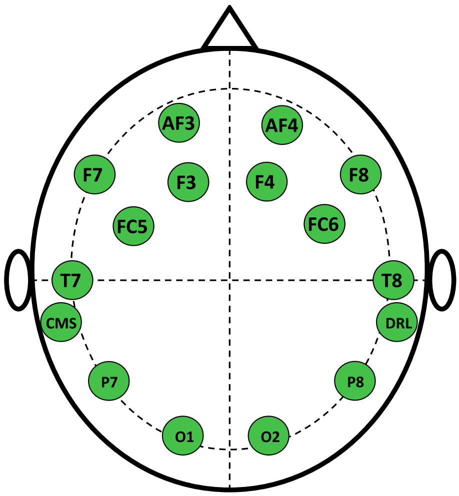


Fig. 2 Stimuli sequence while the EEG data are collected. Initially, the participants are given a blank screen to help with visual stability. Then, at random intervals, a set of stimuli for a certain product are shown through the study (first a product, then its endorsement or promotion)

or 50% off with that particular product. Figure 3 demonstrates the stimuli where each row is for different products. The products were burger, sun-glass, cake, hat, and coat. Here, for each product, the first column is the baseline product, second column is for endorsement, and the last column is for promotion stimuli (Fig. 4).

Fig. 3 All stimuli with corresponding experimental setup where the first column is the baseline product, second column is for endorsement, and the last column is for promotion stimuli

Fig. 4 Emotiv EPOC+ electrode positioning



2.3 Data Collection

The data collection process is separated into three stages. In stage 1, the experimenter briefs the participants about the stimuli so that they will be at ease when watching them on screen. In the second stage, participants sit in front of a monitor that displays the stimuli at a 75–100 cm distance. While the participants are viewing the stimuli, EEG data are acquired from them using an Emotiv Epoch+ headset which is a wearable headset that contains 14 channels according to the 10–20 electrode system. The sampling rate 128 Hz. For this work, we use six frontal channels as they showed better performance in previous studies [12]. Illustrated in Fig. 2, each product has been shown followed by its endorsement for five seconds throughout the experiment. Note that, for each product, we randomize stimuli. In stage 3, we gave the participants a questionnaire where each stimulus was presented with the statement: 1. I would feel happy to buy x and 2. Given the chance, I am willing to buy x. Subjects answered on a numeric scale of 1–10 (strongly disagree to strongly agree) that has later been converted to a binary form: negative (1–5) and positive(6–10).

2.4 Pre-processing of EEG Signals

The collected EEG signals are pre-processed and analyzed using MATLAB 2020a (MathWorks, Natick, MA) software. We also use the customized script along with EEGLAB. Initially, we extract the EEG signals between 0.5 48 Hz using a sixth-order bandpass Butterworth filter, followed by 50 Hz notch filter for removing the power line noise. The contamination of eye artifacts, line noise, and movement artifacts are then removed using independent component analysis. Lastly, we segment the signal with only stimuli viewing five seconds.

2.5 Wavelet Packet Transform (WPT)

The signals are decomposed by WPT into both detailed and approximation coefficients. The extracted coefficients by WPT up to a specific level could be termed EEG signal features because the value of the coefficients is dependent on the time and frequency domain characteristics of the EEG signal. WPT creates a subspace tree of a signal with various frequency characteristics by recursively applying highpass and lowpass filters [8]. Let, $\prod_{a,b}(k)$, $n = 0, \ldots, 2^a - 1$, denote the WPT coefficients at level a. The two wavelet packet orthogonal bases equations are then used to calculate the wavelet packet coefficients:

$$\prod\nolimits_{a,2b}(i) = \sum_{l=0}^{L-1} H(s) \prod\nolimits_{a-1,b} (2k + 1 - \text{lmod}\, N_{b-1}) \tag{1}$$

$$\prod_{a,2b+1}(i) = \sum_{l=0}^{L-1} G(s) \prod_{a-1,b} (2k + 1 - l \bmod N_{b-1}) \tag{2}$$

where $H(s)$ and $G(s)$ are the impulse responses which are highpass and lowpass filters of the wavelet packets, respectively, and $i = 1 \ldots N$ and $N_b = N/2_b$ [8]. In this work, we decompose the signals into six bands, namely $\delta = 0 - 4$Hz, $\theta = 4 - 8$Hz, $\alpha = 8 - 12$Hz, $\beta_1 = 12 - 20$Hz, $\beta_2 = 20 - 32$Hz, $\gamma = 32 - 64$Hz.

2.6 Feature Extraction

For any participant Y_N, segmented time series vector for one electrode is $X(t) \in \mathbb{R}^T$, where T represents the number of samples in time. We extract three types of features from the EEG signals, namely time domain, frequency domain, and time–frequency domain features.

- Average power (ψ): $\psi = \frac{1}{N} \sum_{i=1}^{N} |X(i)|^2$
- Relative power (ζ): $\zeta = \frac{\psi}{\rho}$, where ρ is the total power of y_t. We also calculated all possible combination of distinct ratios ($\frac{\delta}{\theta}, \frac{\delta}{\alpha}, \frac{\delta}{\beta_1}, \frac{\delta}{\beta_1}, \frac{\delta}{\gamma}, \frac{\theta}{\alpha}, \frac{\theta}{\beta_1}, \frac{\theta}{\beta_2}, \frac{\theta}{\gamma}, \frac{\alpha}{\beta_1}, \frac{\alpha}{\beta_1}, \frac{\alpha}{\gamma}, \frac{\beta_1}{\beta_2}, \frac{\beta_1}{\gamma}, \frac{\beta_2}{\gamma}$) for both ψ and ζ.
- Arithmetic mean (μ): $\mu = \frac{1}{N} \sum_{i=1}^{N} X(i)$
- Modified mean absolute value (MMAV):

$$\text{MMAV} = \frac{1}{N} \sum_{i=1}^{N} w_i \, |X(i)|;$$
$$w_i = \begin{cases} 1, & \text{if } 0.25\,N \leqslant i \leqslant 0.75\,N \\ 4i/N, & \text{elseif } i < 0.25\,N \\ 4(i-N)/N, & \text{otherwise} \end{cases}$$

- Standard deviation (λ):

$$\lambda = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(X(i) - \frac{1}{N} \sum_{i=1}^{N} X_i \right)^2}$$

- Skewness (S): $S = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\lambda} \right)^3$
- Kurtosis (κ): $\kappa = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\lambda} \right)^4$

- Median (ν): $\nu = X(\frac{i}{2})^{th}$
- Difference variance value (DVV):

$$\text{DVARV} = \frac{1}{N-2}\sum_{i=1}^{N-1}(x(i+1) - x(i))^2$$

- Difference of absolute standard deviation (DASD):

$$\text{DASD} = \sqrt{\frac{\sum_{i=1}^{N}(X_{i+1} - X_i)^2}{N-1}}$$

- Mean Teager value (MTE):

$$\text{MTE} = \frac{1}{N}\sum_{i=k-N+3}^{k}\left(X(i-1)^2 - X[i]X[i-2]\right)$$

- Renyi entropy (RE):

$$\text{RE} = \frac{1}{1-\alpha} \times \log\left(\sum\left(\frac{X^2}{\sum X^2}\right)^{\alpha}\right)$$

  $\alpha \neq 1$ and α value is 2 in this work.
- Teager–Kaiser energy operator (ξ) measures the instantaneous change of energy which is defined as:

$$\xi = \log\left(\sum_{i=1}^{N-1} X^2(i) - X(i-1)X(i+1)\right)$$

- Normalized 1st and 2nd difference (χ_1 and χ_2) is calculated by:

$$\chi_1 = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|X(i+1) - X_i|}{\lambda}$$

$$\chi_2 = \frac{\frac{1}{N-2}\sum_{i=1}^{N-2}|X(i+2) - X_i|}{\lambda}$$

- Hjorth features

$$\text{Activity} = \frac{\sum_{t=1}^{T}(X(i) - \mu)^2}{T}$$

$$\text{Mobility} = \sqrt{\frac{X\dot{(i)}}{(X(i))}}$$

$$\text{Complexity} = \frac{M(\dot{X}(i))}{M(X(i))}$$

- Log coefficient variation (LCV):

$$\text{LCV} = \log\left(\frac{\sqrt{\frac{1}{(N-1)}\sum_{i=0}^{N-1}(X[i] - \dot{X})^2}}{\frac{1}{N}\sum_{i=0}^{N-1}X[i]}\right)$$

- Temporal moment (TM) is calculated in 3rd order which is expressed by: $TM = \left|\frac{1}{N}\sum_{i=1}^{N}X(i)^3\right|$
- Shannon entropy (SE): $SE = -\sum_{i=1}^{N}\frac{X^2}{\sum_{i=1}^{N}X^2}\log_2\frac{X^2}{\sum_{i=1}^{N}X^2}$
- Threshold crossing (TC) is the number of times that amplitude values cross zero amplitude from a threshold value $T = 4 \times \frac{1}{10}\sum_{i=1}^{10}X(i)$ which iterated $N - 1$ times. Each time the below condition satisfies, the value of TC increases by 1.

$$TC = \begin{cases} 1, & \text{if X}(i) > T \text{ and X}(i+1) < T \\ & \text{or X}(i)\text{i} < T \text{ and X}(i+1) > T, \\ 0, & \text{otherwise.} \end{cases}$$

Time-domain features are calculated from $X(t)$ while decomposed in the temporal domain. All the aforementioned features are extracted in the time domain. On the other hand, the frequency domain features are extracted from the spectral features of the signal. We extract centroid [7], spread [7], kurtosis [7], entropy [6], decrease [7], roll off point [9] in spectral domain. Despite their ubiquitous use in speech and audio signal classification, these features have just recently been used for EEG signal categorization. Spectral features capture the amplitude spectrum of EEG data, which give discriminatory information between classes.

EEG signals are sophisticated in nature as they have qualities in both the temporal and frequency domains. EEG signals are split in this work utilizing wavelet packet transformation (WPT), which may recover frequency information without leaving the temporal domain. In the literature, WPT has been routinely utilized to distinguish frequency bands from EEG signals [10, 11].

2.7 Feature Selection and Classification

To select the best feature set, wrapper-based support vector machine-Recursive Feature Elimination (SVM-RFE) along with correlation bias reduction (CBR) is implemented [13]. In this study, we performed three evaluation schemes, namely k-fold cross-validation, where k=3. We use SVM for the classification of positive and negative affective attitude (AA) and purchase intention (PI). SVM utilizes a kernel function to transform the inputs into a high-dimensional feature space, whereas learning derives the decision boundaries directly from the training dataset. It then constructs an optimal separation hyperplane in the feature space. Four different kernels are used, namely linear, polynomial, radial basis function (RBF), and sigmoid. To improve classification performance, the hyper-parameters of an SVM classifier, namely the regularization parameter C and the γ, are optimized during training with a range 2^{10} to 2^{-10}. We use LIBSVM [2] function for classification of our work (Fig. 5).

3 Result and Discussion

This section presents the cross-validation results obtained by the proposed model. We also compare the different kernel results for both AA and PI analysis along with the discussion of the performance.

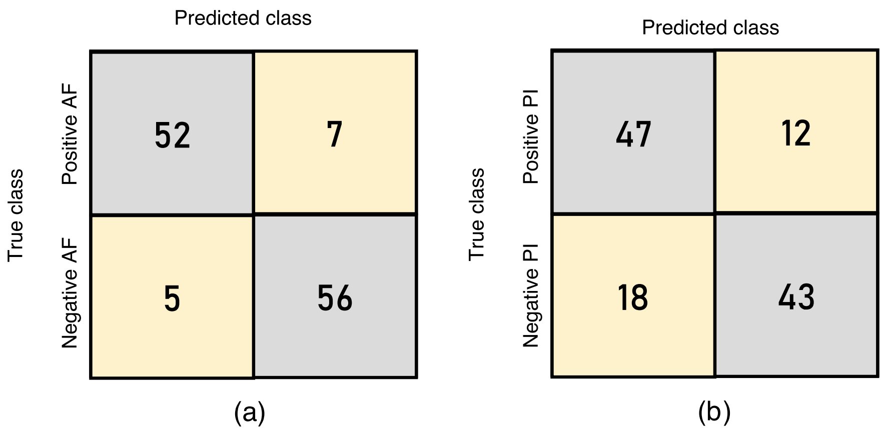


Fig. 5 Confusion matrix of the analysis. **a** Affective attitude, **b** purchase intention. Results are shown as combination of threefold cross-validation

3.1 Affective Attitude Analysis

Based on the extent literature, it appears that this work presents the first marketing stimuli-based ML classification framework. In Table 1 reports the performance of the best achieved AA classification results of accuracy 90 ± 4.33, sensitivity 88.15 ± 7.5, and specificity 91.74 ± 3.02. This performance is superior to existing neuromarketing related work [12], though they did not use directly any marketing stimuli. We also compared our results with four kernels of SVM, namely, linear, polynomial, RBF, and sigmoid. Figure 6a illustrates that the RBF kernel outperforms the other solving the problem of spatial complexity. Note that, to understand the model's stability, we perform cross-validation 20 times for each classifier which has been shown as box plot in Fig. 6.

3.2 Purchase Intention Analysis

Based on the extent literature, it appears that for the first time, this work proposed a PI-based ML framework. Table 1 reports the performance of the best achieved PI classification results of accuracy 75 ± 2.5, sensitivity 79.65 ± 8.12 , and specificity

Table 1 Performance using SVM classifier

Outcome	Accuracy	Sensitivity	Specificity
Affective attitude	90 ± 4.33	88.15 ± 7.5	91.74 ± 3.02
Purchase intention	75 ± 2.5	79.65 ± 8.12	70.55 ± 2.68

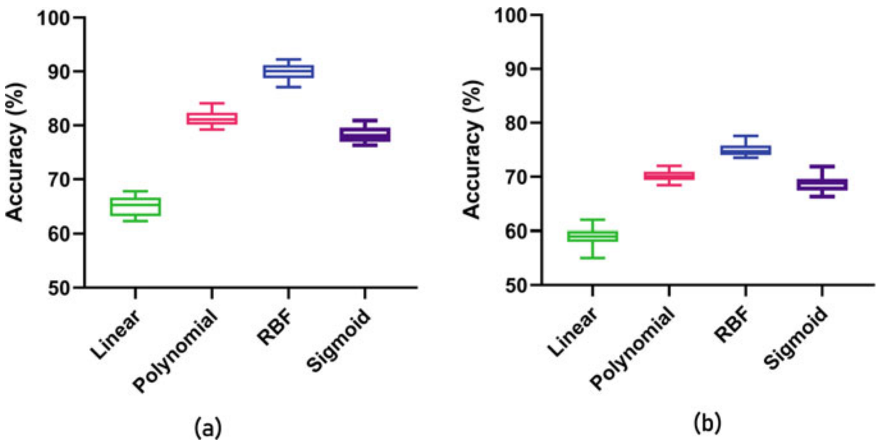


Fig. 6 Performance of the proposed framework for various kernels of SVM shown in box plot. **a** AA, **b** PI and SVM-RBF outperforms other kernels

70.55 ± 2.68. Similar to AA, Fig. 6b four kernel performance using SVM classifier where the performance of the linear classifier was poor compared to the nonlinear classifiers. Similar to previous results, RBF performs best in this case.

In this work, we use wrapper-based feature selection technique which uses SVM ML model while selecting the best features. That is why we used SVM as the classifier. In addition, we also use other models such as naive Bayes, decision tree , k-Nearest Neighbors which yields an accuracy of 66.66%, 67.34%, and 71.12% for AA and 54.72%, 56.33%, and 59.89% for PI, respectively. It can be seen that the results are significantly lower than SVM; this is due to the fact that the features are selected in such a way that these give the best performance using SVM.

One of the limitations of this work is number of subjects that we can increase in future studies to improve the generalizability of the work. Participants in this study are limited to young adults subjects considering them as target consumers of the marketing stimuli; in the future, a more diverse subject group should be included alongside different intervals of purchase like daily required products, weekly or monthly, and product-groups like fresh, stationery, home or office appliances, etc. In addition, future work might consider comparing EEG results with other modalities.

4 Conclusion

In this work, we established a system to predict customer choice in marketing stimuli using EEG data using ML approach. The findings indicate that the proposed strategy is effective and may be used to supplement traditional methods for predicting client responses to marketing stimuli. We also showed that RBF kernel performs the best with the SVM classifier. Finally, it is clear that neuromarketing technology is effective in forecasting customer behavior and predicting future consumer preferences.

Acknowledgements This study was funded by IAR, UIU (Code: IAR/01/19/SE/10) and Market-Brain: A Neuromarketing System for Advanced Market Research Project Under ICT innovation Fund, ICT Division, MoPTIT, GOB, Project Code No. 1280101-120008431-3631108.

References

1. Rawnaque, F.S., et al.: Technological advancements and opportunities in neuromarketing: a systematic review. Brain Inf. **7**(1), 1–19 (2020)
2. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 1–27 (2011)
3. Chew, L.H., Teo, J., Mountstephens, J.: Aesthetic preference recognition of 3D shapes using EEG. Cogn. Neurodyn. **10**(2), 165–173 (2016)
4. Doborjeh, Z.G., Doborjeh, M.G., Kasabov, N.: Attentional bias pattern recognition in spiking neural networks from spatio-temporal EEG data. Cogn. Comput. **10**(1), 35–48 (2018)

5. Jin, J., Zhang, W., Chen, M.: How consumers are affected by product descriptions in online shopping: event-related potentials evidence of the attribute framing effect. Neurosci. Res. **125**, 21–28 (2017)
6. Misra, H., Ikbal, S., Bourlard, H., Hermansky, H.: Spectral entropy based feature for robust ASR. In: 2004 IEEE ICASSP, vol. 1, pp. I–193, IEEE (2004)
7. Peeters, G.: A large set of audio features for sound description in the CUIDADO project. CUIDADO Project Rep. **54**, 1–25 (2004)
8. Percival, D.B., Walden, A.T.: Wavelet Methods for Time Series Analysis, vol. 4. Cambridge University Press (2000)
9. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: IEEE ICSSP, vol. 2, pp. 1331–1334, IEEE (1997)
10. Vidyaratne, L.S., Iftekharuddin, K.M.: Real-time epileptic seizure detection using EEG. IEEE Trans. Neural Syst. Rehab. Eng. **25**(11), 2146–2156 (2017)
11. Wali, M.K., Murugappan, M., Ahmmad, B.: Wavelet packet transform based driver distraction level classification using EEG. In: Mathematical Problems in Engineering (2013)
12. Yadava, M.E.A.: Analysis of EEG signals and its application to neuromarketing. Multimedia Tools Appli. **76**(18), 19087–19111 (2017)
13. Yan, K., Zhang, D.: Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors Actuators B: Chem. **212**, 353–363 (2015)

An Empirical Analysis of IT-Software Job Skill Requirements During COVID-19 Pandemic Period in Bangladesh

Mijanur Rahaman, Md.Masudul Islam, and Md.Saifur Rahman

Abstract This paper mainly focuses on the job requirements of the IT-Software industry during the COVID-19 pandemic situation using recent data on vacancy posting and job ads views on Bangladesh's largest online job portal. The impact of COVID-19 in every sector in Bangladesh had started in March 2020. This pandemic made us adopt the online system almost in every aspect of our life. That is why the IT-Software job industry and the requirements in these jobs have reformed newly. This paper will show the major requirements and skills for the IT-Software job during this pandemic situation and help us to understand which skills we should develop in this pandemic situation.

Keywords COVID-19 · Coronavirus · Job · IT-Software · Requirements · Skills · Demand · Vacancies · Online job portal · Digital Bangladesh

1 Introduction

COVID-19 is the name of the epidemic that becomes a pandemic by affecting globally these days. It has a greater negative impact on our worldwide economic, social, health, and educational activities. Particularly, the job markets are impacted by unemployment, and job postings are decreasing. The intensity of job searching is increasing but there is a huge gap between employment and job search in recent years. In the case of Bangladesh, one of the leading development country unemployment in some job sectors is alarming. During this pandemic situation, people become domesticated in their homes. Two major concepts grow up in this pandemic; one is keeping social distance to prevent spreading viruses and another one is working from home to

M. Rahaman · M. Islam (✉) · M. Rahman
Bangladesh University of Business and Technology (BUBT), Dhaka, Bangladesh
e-mail: masud@bubt.edu.bd

M. Rahaman
e-mail: riponcse.it@bubt.edu.bd

M. Rahman
e-mail: saifurs@bubt.edu.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_4

continue earning the livelihood even though it is not much effective. That is why the world is now adopting the online system in their day-to-day life as if it is becoming our sixth fundamental need. In this context, the growth of IT-Software industries is creating a vast opportunity for other job fields and holding them from totally extinct. A good example would be our worldwide online education system and E-learning during the COVID-19 period that is available for everyone from any corner of the world if and only if they have some facility of digital accessories and Internet. COVID-19 shows us that how important it is to implement a digital system in our social, business, educational, and miscellaneous life that reduces our economic crisis to some extent. This paper will show some background history of the situation of the worldwide job market especially the IT sector. We have summarized COVID-19 impact on worldwide employment more specifically the IT-Software job sector from March 2020 to June 2021. In there we show the loss of employment, job searching activity, potential job market. Also, we will show a summary of the employment condition in Bangladesh and the present state of IT-Software industries. This paper will provide empirical evidence about the IT-Software job requirements or skills during this COVID-19 crisis. We investigate the IT-Software employer's perspective and their given skill requirements from their job post. We analyze real-time data from bdjobs.com, the largest online job portal in Bangladesh. The statistical view of job skill requirements for IT-Software industries could be a guideline for potential learners, students, unemployed people to understand what they should learn or develop in themselves to get a better job opportunity. This paper proceeds as follows. We have discussed a summary of worldwide employment in any sector and then emphasize on worldwide IT-Software job market and then we introduce the employment status of Bangladesh and focusses on the IT-Software employment sector in Sect. 2 Background. The data collection, cleaning, and filtering are discussed in Sect. 3 Data. We represent our statistical view of analyzed data in Sect. 4 and discuss the analysis report in Sect. 5. Finally, we conclude our paper in Sect. 6.

2 Overview of Global Employment in COVID-19 Period

COVID-19 pandemic has become a major threat to our worldwide growth of the economy and overall job sector. Many industries like leisure and hospitality, education and health services, manufacturing, construction, transportation system, wholesale trade, etc. faced job loss during this period. In the USA, the healthcare and education sector saw job cuts with a loss of 2.5 million, and the professional and business services sector, which includes lawyers, accountants, engineers, and temporary workers (such as laborers, office clerks, and packagers lost 2.1 million jobs within the first quarter of 2020 [1]. EU had identified that 35% of jobs can be conducted from home by the online system but in the case of the manufacturing sector the labor and the poorly paid workers are mostly in the pain of job risk [2]. A statistical report of March 2020, shows that during this tentative COVID-19 session (2019–2022) in the European Union waste management, social work, public administration, energy supply, etc. will be the least affected sectors and accommodation, food services, manufacturing, real estate, scientific activities, and transportation services will be the most affected sector [3]. According to a report of marketplace.org from last year food services, doctors, construction firms, retail companies, online shopping, and mobile marketing are some sectors that win the race of jobs in the COVID-19 session. On the other hand, the tourist, arts and entertainment, and nursing jobs' sectors are the verges of death [4].

It seems like during the ups and downs of employments in this pandemic some demand for specific skills in some sectors such as health, logistics, IT-Software, and education has increased. The online job portal reveals the worst job market conditions after March 2020. Some key requirements for jobs have changed the nature of work in this pandemic. For example, most of the jobs are advertising "work from home" as a required condition. Also, most of the jobs require lower qualifications in some countries. These jobs are searching for underlying skills such as "communication skill", "teamwork", "good at the internet". In April 2021, the OECD report shows an assessment of the impact of COVID-19 on job and skill requirements by analyzing job posts [5].

COVID-19 have taught our IT industries a great lesson that how to manage any changes and avoid the drawbacks of crisis management using digital system [6]. The impact of COVID-19 in the technology sector is significant. Usage of Hardware and Software sector has increased in a personal level, IT-Network services and usage of equipment has increased insignificant proportion and semiconductor industry are facing black sawn event during this pandemic [7]. Worldwide technical supplies are canceled, so many tech conferences got canceled, people are fearful of losing their jobs. On the other hand, the work from home culture makes the SaaS software development a boon, streaming, and gaming companies are getting the highest benefits in this lockdown period and the IT industry is taking more innovative efforts to cope up with the challenges in web developments [8]. The world business faced almost USD 1 trillion of the estimated amount in 2020. The global IT services industry is going to see a decline in overall revenues by approximately 3–4 % whereas India alone stands to lose more than USD 5 billion [9]. The pandemic has changed the nature of software engineering approaches. It has created a new software development model called software product engineering which only focuses on the customer and market-oriented software solutions. Also, virtual development teams are collaborating within themselves by regular video conferencing and document sharing [10]. Even though the overall job sector is facing difficulties in this pandemic period the demand for software developers is still strong. The unemployment rate in the IT industry has also remained at 4.2%. This is relatively low as compared to other industries and sectors. Also, we need to remember that without IT consultants, companies will find it difficult to ensure the widespread implementation of their digital transformation solutions. That is why companies are hiring a virtual specialist. The job market for software developers has continued to thrive in 2020 [11]. IT outsourcing companies have become more sought after than it ever was. A study conducted by Computer Economics has revealed that large organizations have witnessed a significant increase in their IT spending the figure this year stands at 8.7%, more than 2% higher than

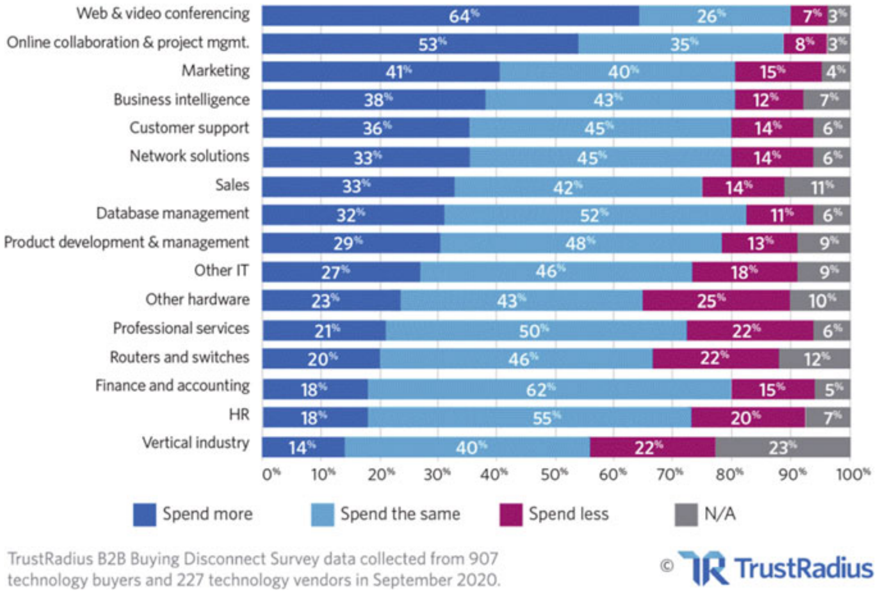


Fig. 1 trustradius.com's survey in 2020

the previous year [12]. According to a survey from 2020 analysis by trustradius.com, most of the buyers expect to spend more on video and web conferencing software (64%) and online collaboration and project management software (53%) in 2021 [13]. In Fig. 1 they showed where the tech buyers will increase and decrease spending in 2021.

A recent Global IT-Software Market Procurement Intelligence Report of SpendEdge forecasts that the spend of the IT-Software market will grow over USD 634.96 billion between the years 2021 and 2025 [14]. We can say the COVID-19 pandemic has changed the way of IT-Software market and the working methodologies by digital transformation.

3 ICT Sector of Bangladesh in COVID-19 Period

Bangladesh which has lower-middle-income is one of the worst-hit countries in south Asia for the COVID-19 outbreak. In this country, out of the employed population of around 63 million, around 25 million are in wage/salaried employment, over 10 million of whom are in casual employment based on daily wages. Approximately 4.5 million casual laborers are engaged daily in construction, transport, trade, and food and accommodation who were seriously affected [15]. The official unemployment rate in Bangladesh is around 4% and every year 2–2.2 million educated unemployed are being added to this list [16]. There are two types of job loss in Bangladesh one is temporary and the other is permanent. In a report of 2020, according to CRI estimates, the "permanent" impact job loss estimates are 6 million, which would nearly double the unemployment rate. It would affect approximately 24 million people [17]. Also, estimates from various sources put the immediate temporary job loss figure (for 2 months of lockdown) between 12 and 17 million [17]. The highest job loss has occurred in agriculture, retail trade, hotels and restaurants, inland transport, construction, textiles and textile products, and other services. Recent research by a2i published that over 20.4 million workers are currently unemployed across 11 high-impact sectors of the economy [18]. The ICT sector of Bangladesh lost USD 1.3 billion of business due to COVID-19 [19]. At present we have three challenges in the ICT sector those are proper financing, skilled manpower, and the high cost of the Internet. The encouraging matter is that the demand for local software is increasing day by day. There is also a light of hope that by the end of 2021 five emerging sectors will generate 40 lakh new jobs said to a report of a2i. According to the head of the future work lab of a2i Asad-Uz-Zaman, the ICT sector will require more e-commerce product managers, data, or cloud engineers while the health sector will need more nurses, medical technicians, and online consultants. The tentative demand for skilled manpower would be 10 lakhs in agro-farming, 10 lakhs in the health sector, 4.5 lakh in pharmaceutical, 6.5 lakh in ICT, and 90,000 in creative media by 2021 [20]. The co-founder of Brainstation-23 said that video streaming, machine learning, and online learning are strong business vertices to explore now. Also, some e-services are getting far better than some others such as Uber, Pathao, and Food Delivery services are falling a bit on the other side Chaldal ordering and some e-health services has increased in a significant manner [21]. So, there is a huge scope of training in these sectors to regain the ICT market because in this COVID-19 and post COVID-19 period the world will become hyper-digital. Cloud, Mobile, social network, and big data analysis will become a base for world IT systems. Soon everyone will feel the necessity of IoT, cloud computing, network management, online services, blockchain, AI, AR, robotics, e-learning, and e-health service in day-to-day life. To take lead in the ICT sector with more than 2000 software companies and 500,000 people skilled in IT, it is time for the industry and the Bangladesh government to sit together and create a well-planned roadmap [22].

4 Research and Methodology

This paper shows a quantitative statistical analysis of skill requirements for the IT-Software job sector in Bangladesh during the recent COVID-19 period. Data is collected from bdjobs.com one of the largest and famous online job websites in Bangladesh. We have collected a total of 358 authentic data from bdjobs.com in the IT and Engineering category using web crawling. These job posts are from March 2021. There are also many online job portals, but we have taken only one and the largest one to reduce redundancy and complexity of cleaning and filtering the same job. Figure 2 shows some steps of how we processed our data. We took the raw data

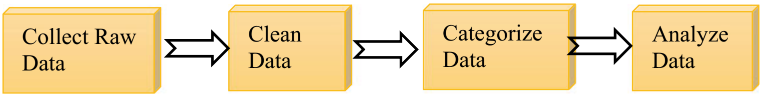


Fig. 2 Data collection and processing

and then clean them by removing unwanted items, text, or semantics. After that, we categorized the data according to job skills. Finally, we analyzed our numerical data in various statistical formats.

5 Skill Requirements Analysis in IT-Software Job During COVID-19 Pandemic Period

The potential of the IT sector in COVID-19 is blooming. To keep pace with the recent situation, we need to find out the updated requirements of IT-Software job fields in our country. We have analyzed some skill requirements in the IT-Software industry. The survey contains 2021s data from the country's leading job site bdjobs.com. We searched for IT/Engineering category jobs and among them, we took 358 valid job circulars around the country. From there we have collected requirements and additional skill requirements concerned with IT-Software. The following skills in Table 1 we fetched 34 types of skills from there.

The following charts will show different scenarios of skill requirements in the IT-Software job sector. In Fig. 3 the chart shows the IT-related jobs according to appear along with categories. According to the chart, 40.22% of the total job asked for fluent, interactive, and linguistic communication skills. Almost 28.21% of the demand for scripting languages like JavaScript, jQuery, Angular, Vue, etc. The demand for Database management and Query analysis skill is 22.07%, project management skill is 21.23% and front-end skills such as HTML-CSS is 19.27%. These are the top 5 skills that are asked mostly in IT-Software jobs. Another analysis of job data is finding the frequency of skill requirements in each job according to their total number of requirements. From our 358 data we have found at least 1 and at most 13 types of skill requirements for each job. In Fig. 4 we are showing the overall frequency of skill requirements in each job using a range of 1–13. We have calculated the weight of every skill based on the number of requirements asked in each job and the specific skills required in jobs. The formula we developed to find this result is:

$$R_w = \left(\sum_{i=1}^{n} \mathrm{Rc}_i * S_i\right) / \sum_{i=1}^{n} \mathrm{Rc}_i$$

Here,

R_w Each Requirements' weight

Table 1 34 types of skills according to their category

Category	Specific skills
Basic IT and tools	Basic IT/Computing Knowledge
	Management Tools (MS Project/Jira/Teamwork)
	Hardware & Software Maintenance
Database	DBMS/MS-SQL/MySQL/Oracle/NoSQL/PostrGL
E-commerce	Digital Marketing/IT Sales/SEO
	E-Commerce/E-Business
Mobile and game	Android
	iOS
Networking and server	Networking/Server Management
	LINUX/Red Hat/Debian
	Cloud (AWS/Management)
	Security
	IIS
Others	Communication
	Graphic Design/Animation
Programming	Algorithm/Problem Solving
	OOP
	.NET/ASP.NET/MS-Framework
	Java
	C/C + + /C#
	Python/Django
Software engineering	Project Management/ERP
	Design Patterns
	Software Testing/Quality Assurance
	Documentation/Content Writing
	System Analysis & Design
Web engineering	JSON/JS/NodeJS/Angular/Vue
	HTML/CSS/Design Framework
	PHP
	UX/UI
	Software/Web Engineering/Development
	Web Framework/CMS/API
	Git/Repository/Version control/Docker
	Laravel

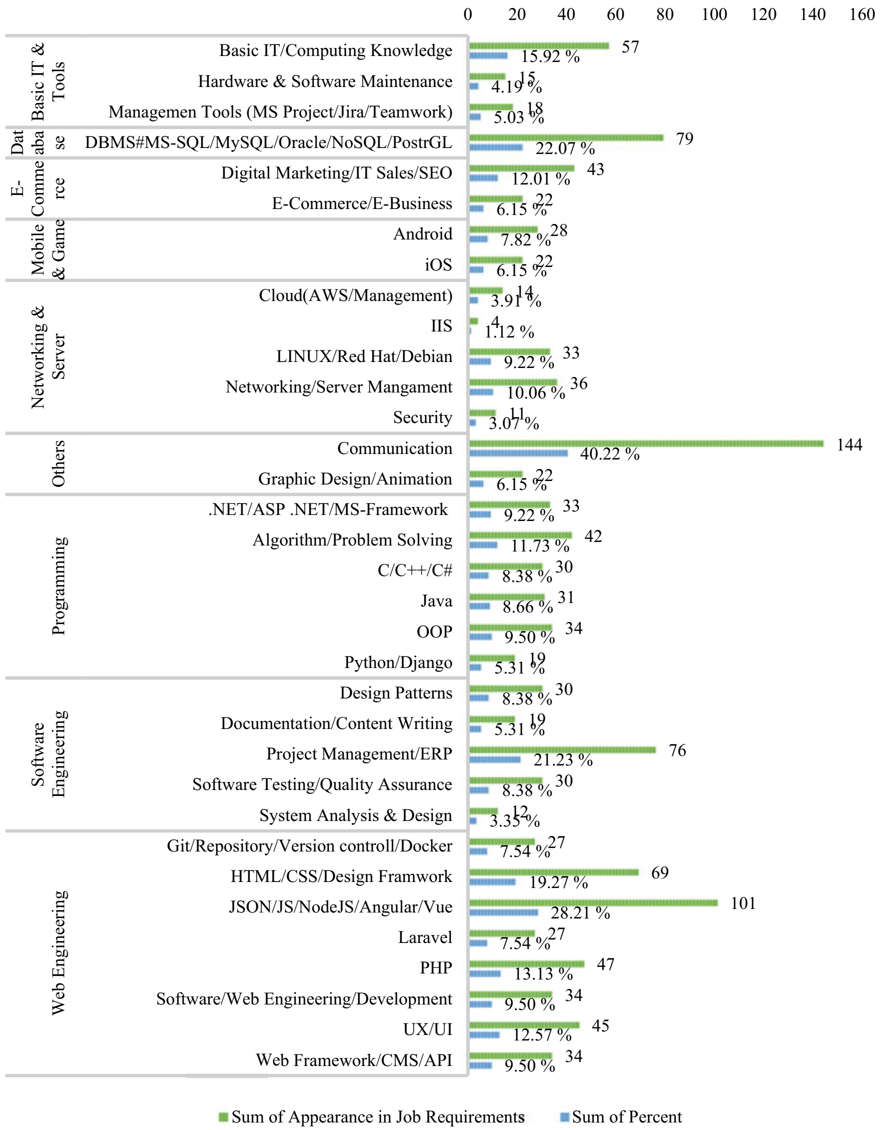


Fig. 3 The appearance of skill (with percentage) requirements in job according to category

Rc Number of requirements counted in each job
S Individual Skill count (frequency) in each job within the range of Rc
n Maximum Number of requirements count in each job.

For example, according to our data, $n = 13$ and there is 27 job post which has asked for 1 type of skill as requirements. Among them, 3 job post has asked for

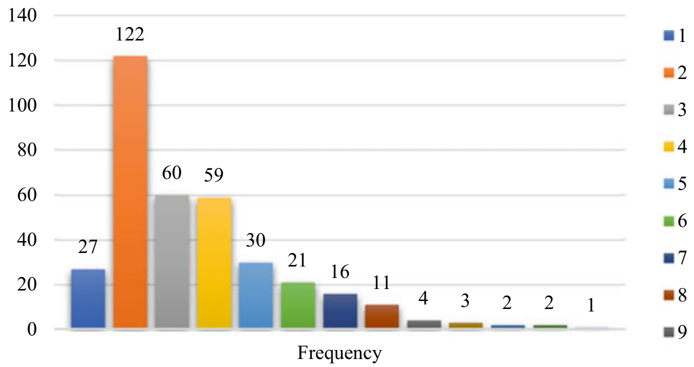


Fig. 4 Frequency of skill requirements in each job

communication as a skill out of 1 requirement. Similarly, there is 122 job post which has asked for 2 skills and among them, 42 of them asked for communication. Therefore, we can see that there is 60 job post that asked 3 types of skills and 28 of them are communication, 59 job post that asked 4 types of skills and 24 of them are communication, 30 job post that asked 5 types of skills and 12 of them are communication, 21 job post that asked 6 types of skills and 12 of them are communication, 16 job post that asked 7 types of skills and 10 of them are communication, 11 job post that asked 8 types of skills and 5 of them are communication, 4 job post that asked 9 types of skills and 4 of them are communication, 3 job post that asked 10 types of skills and 1 of them are communication, 2 job post that asked 11 types of skills and 1 of them are communication, 2 job post that asked 12 types of skills and 2 of them are communication, 1 job post that asked 13 types of skills and 0 of them are communication. Therefore, our weight for communication skills is

$$
\begin{aligned}
R_w(\text{Communication}) &= (1 \times 3 + 2 \times 42 + 3 \times 28 + 4 \times 24 + 5 \times 12 \\
&\quad + 6 \times 12 + 7 \times 10 + 8 \times 5 + 9 \times 4 + 10 \times 1 \\
&\quad + 11 \times 1 + 12 \times 2 + 13 \times 0)/(1 + 2 + 3 + 4 \\
&\quad + 5 + 6 + 7 + 8 + 9 + 10 + 11 + 12 + 13) \\
R_w(\text{Communication}) &= 590/91 \\
R_w(\text{Communication}) &= 6.48
\end{aligned}
$$

This is how we have calculated weight for every single skill. In Fig. 5, we have shown the overall weighted graph chart (high to low).

Our analysis and graph chart shows different types of scenarios based on different aspects in IT-Software job sector skill requirements. From this analysis, we will clarify some thoughts and recommend some ideas to emphasize on developing major skills in the IT-Software field.

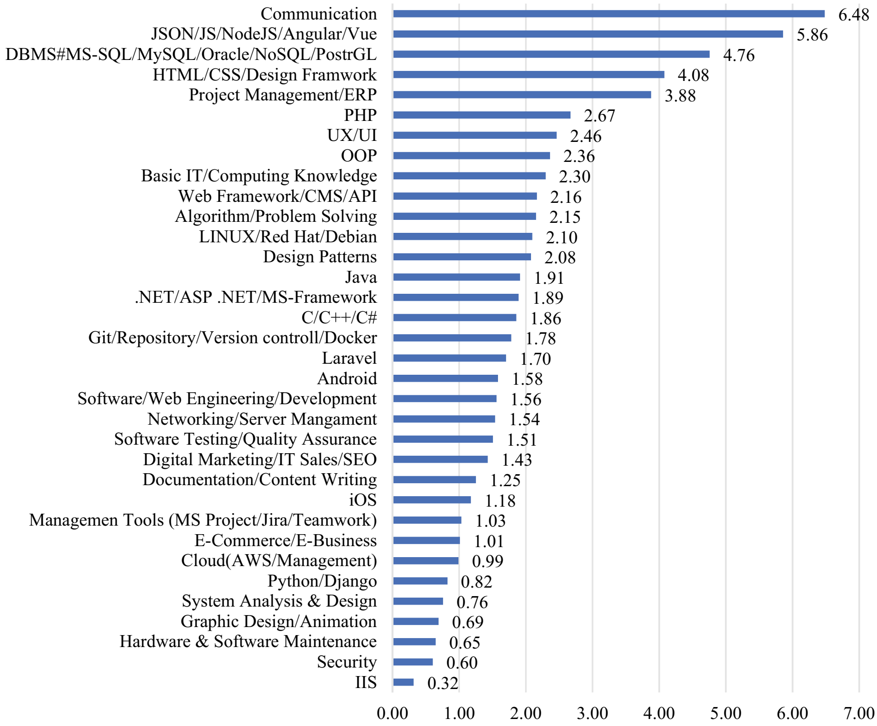


Fig. 5 Weighted graph for individual skill requirements

6 Discussion and Recommendation

For the last fifteen years, universities around the world are continuously developing effective curricula for Software and Web Engineering to create good opportunities for graduates to cope up with IT-Software industries. We can easily find some major gaps between the skill requirements of IT-Software industries and universities' course curricula in real life. In a study of 2020, Mijanur Rahaman & Md. Masudul Islam has shown the gap between the skill requirements of IT-Software industries and universities' web course curricula. They also suggested a well enough balanced curriculum for web courses for university [23]. We need to change our conventional way of studying for IT-Software field and need more interaction with industries to reduce the practical gap, especially in the COVID-19 period. From our study, we present an overall view of IT-Software industries' job status in this COVID-19 situation.

From our initial analysis chart in Fig. 3, we can see that, top five required skills in IT-Software job post are (high to low) Communication (144), Scripting languages such as JS, JSON, jQuery, Angular, Vue, Node JS, (101), Database Management and Query Language (79), Project Management & ERP (76), and Front-End such as

HTML, CSS, Web Design framework, (69). In the percentage scenario, these five has also shown the foremost impact among these 358-job post.

We can say that the recent IT-Software job industries are asking mostly Web/Software, Database skill-oriented people who have good communication skill. In Fig. 4, we can see 122 job post has asked for two (2) types of skill requirements. On average, most job posts demand 1–3 types of skills. There is also a handsome amount of job post that has asked for 4–6 types of skills. This shows that the IT-Software job sectors are focusing on what they need most, and they are mostly not randomly asking for the one-man army in this case. Also, we can see that there are some differences between the weighted graph chart in Fig. 5 and the appearance chart in Fig. 3. The top five weighted skills we find here are (high to low) Communication (6.48), Scripting language such as JS, JSON, jQuery, Angular, Vue, Node JS, (5.86), Database Management and Query Language (4.76), Front-End such as HTML, CSS, Web Design framework, (4.08), and Project Management and ERP (3.88). Here, the Front-End skill has left the project management skill behind. From this weighted chart, we can deeply observe the insight of IT-Software job skill requirements. Therefore, it is clear to us that besides the requirements of the Software and Web field the communication skill is an uprising in today's world.

Now it's time to focus our development on the IT-Software field especially in the scripting language, front-end development, database system, project management, and largely in communication fluency. We need to understand that, developing software and having expertise in the IT sector is a must thing but keep up with this competitive job market everyone must be expert in interaction with each other both offline and online. Since in COVID-19 periods everyone is working from home, so the major demand is fluent and excellent communication from a remote area to any client or people. Also, we should add an extra curriculum in our textbook, and we can call it "IT First Aid" where general people and specialists can learn how to deal with Information and Communication Technology during a disastrous situation. This subject can provide emergency and immediate IT knowledge to anyone, and the active learners can also be benefited by developing their basic skills in IT. There should be offline and virtual training programs that can create skilled manpower from time to time based on the demand of IT-Software industries. Finally, to keep up with this COVID-19 job situation the government and private sector need to take immediate action to provide proper training, encouragement, and industrial curriculum in educational institutions. This is how we can at least hold the IT-Software job industries which have a big potential market in front of our eyes.

7 Conclusion

The COVID-19 pandemic has created some unique scenarios in various job sectors, especially in the IT job sector. This paper reports the first study of how IT-Software industries are demanding different types of skills. Furthermore, this study is exceptional because we have shown the recent IT-related job status in the pandemic situation as well as the statistical summary of skills demanded in various job posts in our country. This study will help our present job holder, potential candidate, and other specialists in IT-Software industries to understand what they should focus on to enhance and stable their job status. This overview will also create insights within IT-Software job candidates so that they can achieve the specific skill to contribute IT sector in this pandemic period. We hope that based on our study we can measure the potentiality of various IT-Software job skill requirements from now and take proper steps to stable our IT job field. This is how we can catch a huge opportunity of the global IT market in the COVID-19 pandemic.

Acknowledgements We are grateful to Mizanur Rahman Mahin, Asst. Programmer and Officer of IT department, Bangladesh University of Business and Technology (BUBT) for his assistance in our research study.

References

1. Franck, T.: Hardest-Hit Industries: Nearly Half the Leisure and Hospitality Jobs Were Lost in April. CNBC (2020). www.cnbc.com/2020/05/08/these-industries-suffered-the-biggest-job-losses-in-april-2020.html
2. Sanchez, D.G., Parra, N.G.: Which Jobs are Most at Risk Because of COVID-19? Brookings (2020). www.brookings.edu/blog/future-development/2020/05/18/which-jobs-are-most-at-risk-because-of-covid-19
3. Coronavirus Impact on Jobs in EU Sectors and Occupations: A Skills. CEDEFOP (2021). www.cedefop.europa.eu/en/news-and-press/news/coronavirus-impact-jobs-eu-sectors-and-occupations-skills-forecast-analysis
4. Tong, S.: Which Jobs are Coming Back First? Which May Never Return? Marketplace (2020). www.marketplace.org/2020/06/17/which-jobs-are-coming-back-first-which-may-never-return
5. An Assessment of the Impact of COVID-19 on Job and Skills Demand Using Online Job Vacancy Data. OECD (2021). www.oecd.org/coronavirus/policy-responses/an-assessment-of-the-impact-of-covid-19-on-job-and-skills-demand-using-online-job-vacancy-data-20fff09e/#section-d1e88
6. Navarro, J.: 5 Things That Software Industry Can Learn from COVID-19. Hexacta—Nearshore Development Software (2020). www.hexacta.com/5-things-that-software-industry-can-learn-from-covid-19
7. Understanding COVID-19's Impact on the Technology Sector. Deloitte (2020). www.deloitte.com/global/en/pages/about-deloitte/articles/covid-19/understanding-covid-19-s-impact-on-the-technology-sector-.html
8. How Will the Coronavirus Impact the Software Industry? (2020). https://customerthink.com/how-will-the-coronavirus-impact-the-software-industry
9. Danley, K.: How COVID-19 Affected the IT/Software Industry Globally. Straits Research (2020). https://straitsresearch.com/blog/how-covid-19-affected-the-itsoftware-industry-globally
10. Software Engineering Takes on New Meaning in the COVID-19 Pandemic. Digitally Cognizant (2020). https://digitally.cognizant.com/software-engineering-takes-on-new-meaning-in-the-covid-19-pandemic-codex5676

11. Tangerine, W.: Is COVID Changing the Job Market for Software Developers? Decide Consulting (2020). https://decideconsulting.com/is-covid-changing-the-job-market-for-software-developers
12. RSK Business Solutions: Impact of Covid19 on IT Software Outsourcing Industry. RSK Business Solutions (2020). www.rsk-bsl.com/blog/impact-of-covid19-on-software-outsourcing-industry
13. How Coronavirus Impacted B2B Software Spending. Trust Radius (2020). www.trustradius.com/vendor-blog/covid-19-tech-spending-data
14. COVID-19 Impact and Recovery Analysis|IT Software Market Procurement Intelligence Report Forecasts Spend Growth of Over USD 634.96 Billion|SpendEdge. Business Wire (2021). www.businesswire.com/news/home/20210129005133/en/COVID-19-Impact-and-Recovery-Analysis-IT-Software-Market-Procurement-Intelligence-Report-Forecasts-Spend-Growth-of-Over-USD-634.96-Billion-SpendEdge
15. COVID-19 Pandemic: An Analysis of the Healthcare, Social and Economic Challenges in Bangladesh. ScienceDirect (2020). www.sciencedirect.com/science/article/pii/S2590061720300727#bb0060
16. Hossain, M.I.: COVID-19 impacts on employment and livelihood of marginal people in Bangladesh: lessons learned and way forward. South Asian Surv. **28**(1), 57–71 (2021). https://doi.org/10.1177/0971523121995072
17. Ahmed, I., Kamal, S.M.: Bangladesh at Work in the Era of COVID-19: Job Creation and Inclusive Growth. The Opinion Pages (2020). https://opinion.bdnews24.com/2020/06/04/bangladesh-at-work-job-creation-and-inclusive-growth-in-the-era-of-covid-19
18. Uddin, M., Munni, S.: Pandemic Wreaks Havoc on Jobs, Pay. The Financial Express (2020). https://thefinancialexpress.com.bd/economy/bangladesh/pandemic-wreaks-havoc-on-jobs-pay-1598068924
19. ICT Industry Lost $1.3 bn of Business Due to Covid-19. Dhaka Tribune (2020). www.dhakatribune.com/business/2020/06/01/ict-industry-lost-1-3bn-of-business-due-to-covid-19
20. The Report, TBS: 5 Sectors Will Generate New Employment in Post-Pandemic Bangladesh: A2i. The Business Standard (2020). www.tbsnews.net/economy/5-sectors-will-generate-new-employment-post-pandemic-bangladesh-a2i-107809
21. Wing, LightCastle Analytics: Bangladesh IT and Digital Sector Tackling COVID-19 Implications. LightCastle Partners (2020). www.lightcastlebd.com/insights/2020/04/bangladesh-it-and-digital-sector-tackling-covid-19-implications
22. Kabir, R.: Rise of Digital Economy: Can Bangladesh Take the Lead? The Daily Star (2021). www.thedailystar.net/business/news/rise-digital-economy-can-bangladesh-take-the-lead-2050749
23. Kar, S., Islam, M.M., Rahaman, M.: State-of-the-art reformation of web programming course curriculum in digital Bangladesh. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **11**(3) (2020). https://doi.org/10.14569/IJACSA.2020.0110323

Predicting Interest of Beginner-Level Engineering Students for Better Career Using Classification Method

Mohammad Aman Ullah, Mohammad Manjur Alam, Saira Akter Sheuli, and Jisrat Alam Mumu

Abstract At the beginning of the university, students often face confusion in career choices. Most of the students cannot understand what field they are interested. There are many existing systems where some researchers have worked with fresh graduates and some with selected tenth or twelfth-grade students to help them get an idea about one career option. However, those research works failed to help them choose a domain of interest from many opportunities in their field. Therefore, this research aims to propose a system for predicting interest of beginner-level engineering students for better career planning at an early age. This analysis gives a chance of finding students interest in four specific demanding fields (web development field, graphics design field, Android development field, and data science field). This research starts with collecting primary sources of data using structured questionnaires. The data are preprocessed for creating a well-formed dataset. Then, the research is done in two parts: statistical analysis (Chi-square test and binary logistic regression analysis) and machine learning analysis (decision tree, SVM, and multinomial logistic regression). In machine learning analysis, appropriate features are selected first, then build the models. The results from both parts are then evaluated using a confusion matrix and choose the best prediction. The research shows noteworthy results with respect to existing works.

Keywords Students · Career · Interest · Field · Predicting

1 Introduction

EDM uses large amounts of student data from educational sectors to find different patterns to help student's decision-making. In present days, technical world competition is increasing rapidly. So, to cope up with this competition, students have to prepare themselves, especially from the beginning of university life. Students often face confusion in career choosing at the beginning of the university. Most of the

M. A. Ullah (✉) · M. M. Alam · S. A. Sheuli · J. A. Mumu
International Islamic University Chittagong, Chittagong 4318, Bangladesh
e-mail: aman_cse@iiuc.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_5

students cannot understand what field they are interested. On the other hand, some students are preplanned about their careers. However, most of the students do not know which career they should go. This confusion results in a great loss of right potential. Some researchers have worked with university students to assess their capabilities and identify their interests. This research will help them know which job role the candidate should be kept in based on his/her performance and other evaluations [1]. Again, some researchers have selected tenth or twelfth-grade students to help them get an idea about one career option through confusion resolution and save from selecting the wrong career [2].

Therefore, this research aims to propose a system for predicting interest of beginner-level engineering students for better career planning at an early age. This analysis gives a chance of finding students' interest in four specific demanding fields [web development field (WDF), graphics design field (GDF), Android development field (ADF), and data science field (DSF)]. This research will help the fresher's or newly admitted engineering students who are not preplanned about their career to take proper decision, in which field they can proceed or which opportunity they can take. The system will be mainly helpful for university students. It will help the students who struggle to choose the right path in building their careers. The analysis will give them the probable idea that helps them think about their opportunities once again, which may decrease the rate of drop out from the universities. This research will benefit both the institutions and the students.

Various machine learning (ML) classification algorithms (Bayesian methods (BMs), multilayer perceptions (MLPs), sequential minimal optimization (SMO), ensemble methods (EMs), decision trees (DTs), SVM, and logistic regression (LR)) are used in the literature for the system mentioned above. In this study, the proposed model used statistical analysis (SA) with Chi-square test (CST), binary logistic regression analysis (BLR), and ML analysis with a DT, SVM, and multinomial logistic regression (MLR). Here, SA is used for making the dataset more efficient using significant variables or attributes.

The remaining paper is structured as follows: Sect. 2 represents the related works. Section 3 represents research objectives; Section 4 is about methodology; Section 5 describes the data analysis; Section 6 includes the results and their description. Section 7 is the conclusion.

2 Related Works

A large number of researchers were done on the prediction of students' career selection issues. The paper by Roy et al. (2018) has provided a career recommender system that used SVM, XGBoost, and DT to recommend the student's future job [1]. A survey by Gorad et al. (2017) on tenth- or twelfth-grade students and subsequent experiments with C5.0 reveals that students select the wrong career due to confusion [2]. Research by Ansari et al. (2017) has proposed a multilevel expert system for career guidance. They have used quantization and nearest neighbor algorithms for finding the relationship between factors [3]. A dataset was developed with socioeconomic conditions, academic performance, and some additional emotional skill parameters by Mishra et al. (2016) and applied Bayesian methods, MLP, SMO, EM, and DT to predict the employability of Master of Computer Applications students [4]. In [5], students' data were collected and applied different algorithms to predict the students' job placement.

Nie et al. (2017) have proposed a data-driven framework to forecast the career choices of students using campus big data. They have used 10 M behavior data to prove the potentiality of this framework and got random forest, which performs better than others [6]. Choe et al. (2020) have tried to understand the career interest of engineering graduate students using the survey data of 249 students. They have applied regression models to predict the career choice of students. This research shows that engineering students require multiple career options to choose a better one [7]. Lakin et al. (2020) tried to find the professional identity of the 186 engineering students and found identity crisis of engineering students once applied the analysis [8].

Al-Dossari et al. (2020) have developed a recommendation system that could help graduate students select their appropriate career based on their skills. With the experiment on 2255 data records with five standard ML algorithms, the XGBoost algorithm has shown the highest accuracy with 70.47% accuracy [9]. Laguador et al. (2020) have evaluated students' internship performance concerning their skills in the graduate program. The study showed a positive relationship between internship performance and attained skills [10]. Park et al. (2020) predicted the features that may affect the career choice of graduate students using SVM and recursive feature elimination. The study found that both mesosystems and macrosystems have affected the newly enrolled students [11]. Fuchs et al. (2020) have applied MLR models to predict students' intention about teaching careers after completing school [12].

Miller et al. (2020) tried to find the relationship between students' enrollment and subsequent career plans. It is found that students having pre-college engineering courses are more likely want to enroll in engineering courses of university regardless of gender or race [13]. Elmasry et al. (2021) have developed a system that could suggest the major using existing data, which helps the students get proper job after graduation and faculty members guide the students properly [14]. Hoff et al. (2021) have predicted that either personality has a positive impact on career outcomes. The study shows that personality development helps the younger in career decisions [15]. VidyaShreeram et al. (2021) applied ML approaches to predict student's careers and got 93% accuracy as its best [16].

3 Research Objectives

The objectives of this research are:

- To find out the student's field of interests from their feedback.
- To propose a model that enhanced the predictive accuracy of career planning.

4 Methodology

The workflow diagram of the experiment design is shown in Fig. 1. The workflow starts with the collection of data from beginner-level university students. The data are then preprocessed to get a well-formed dataset. The analysis is then sub-divided into statistical and ML experiment analysis. The statistical analysis is done using CST and BRL, whereas for the ML experiment, features were selected first, and then, different ML classifiers such as DT, SVM, and MLR were used for classifying the dataset. Confusion matrix is used to calculate the overall performance evaluation. Finally, results from both types of analysis were compared to reach a final decision. The SA was done using the IBM SPSS statistics data editor (version 25) and ML analysis using the R programming language. The detail of each step of Fig. 1 is described below:

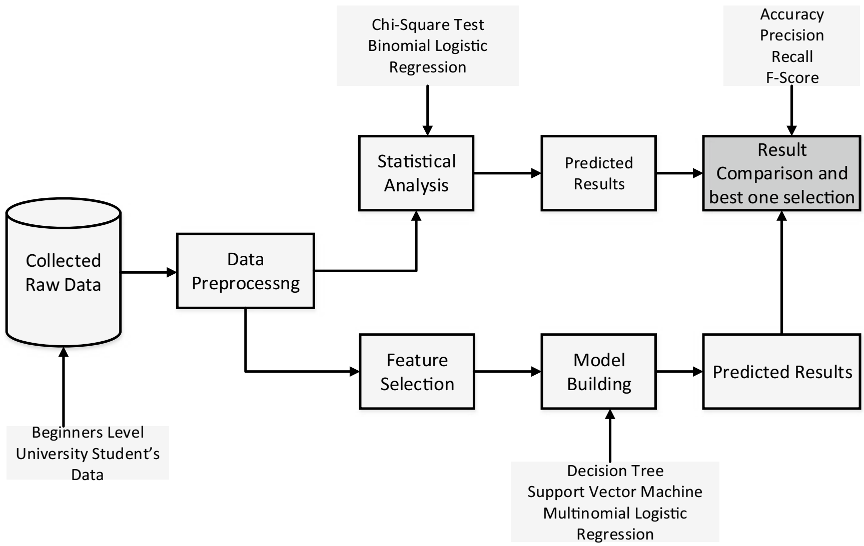


Fig. 1 Workflow diagram of experiment design

Table 1 Reliability and validity of the data

Factors	Number of items	CA if item deleted	CR
WDF	7	0.854	0.702
ADF	7	0.855	0.820
GDF	7	0.849	0.880
DSF	7	0.856	0.817
Overall	28	0.860	

4.1 Data Collection and Preprocessing

In this research, the primary data have been collected from the beginner-level university students of the Department of CSE at International Islamic University Chittagong (IIUC). Data were collected through structured questionnaires on four major career categories in computer science and engineering fields (WDF, ADF, GDF, and DSF). Each field contained ten questions. A total of 250 respondents were randomly selected from CSE students in the 1st, 2nd, and 3rd semesters of spring 2020. After receiving the questionnaires, we found the response of 83 from 1st semester students of different sections, 83 from 2nd semester, and 84 from the 3rd semester.

4.2 Reliability and Validity of the Data

Dataset was prepared by cleaning or removing null and missing data and replacing them with appropriate data. The anyNA() function was used to check inconsistency or null value. Cronbach's alpha (CA) and composite reliability (CR) are most commonly used to assess the measures of reliability and validity of the data. Table 1 shows that the overall reliability coefficient of WDF, ADF, GDF, and DSF is 0.860 that indicates reliability of data as per Hair et al. [17]. All the values of CA if the item deleted are above 0.80, and CR above 0.7 indicates this study that seems high reliable [17].

4.3 Feature Selection Process

Since our dataset contains several features, it was necessary to find the most important features that improve the prediction. Among different types of feature selection processes, we have applied the CST along with CARET package from the RStudio and selected a total of 22 features from our dataset.

4.4 Reading, Training, and Testing Data

In this step, the preprocessed dataset, which is saved in the Haven library of RStudio, is read, trained and tested. The dataset is partitioned into training and testing sets by maintaining 80% and 20% ratios, respectively. After the partitioning, we have found 203 observations for training and 47 observations for testing. The model accuracy and reliability are determined by selecting model with maximum correct prediction.

4.5 Machine Learning Algorithms and Statistical Approach

Data analysis has been accomplished on both ML algorithms and statistical methods. As a statistical approach—CST of independence, BLR analyses were performed. CST is done by Crosstabs procedure. As ML analysis, DT, SVM, and MLR have been used for predicting interest of beginner-level engineering students.

Statistical Analysis of the Study

Chi-Square Analysis. A Chi-square (χ^2) test was used to measure how expectations compare to actual observed data (or model results). The formula for Chi-square is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

where c = degree of freedom, O = observed values (s), and E = expected value (s).

We choose significance levels less than 0.05. In SPSS, the CST is done by the Crosstabs procedure. This procedure creates a two-way table, which summarizes the distribution of categorical variables. From that table, the values of χ^2 test statistic along with p values are found. Since the p-value is less than our chosen significance level (0.05), we can say that all independent variables are highly significantly associated with the dependent variable. Since p-value is greater than our chosen significance level (0.05), there is no statistically significant association between variables.

Binary Logistic Regression Analysis. The model for LR analysis assumes that the response variable, Y, is a categorical variable that wants to be a web developer, Android developer, graphic designer, and data scientist (0 = no and 1 = yes).

Generally, the multiple regression models use the following equations:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2}$$

where Y is the response variable, and the x's are the different predictor variables.

As the response variable of LR is categorical, the outcome variable Y needs to be logit transformed so that the following regression process can be used.

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{3}$$

where P represents the probability of an event occurring, and $\left(\frac{P}{1-P}\right)$ measures the odds ratio. In this study, the BLR model is used to evaluate and identify the effect of variables on student interest achievement.

Machine Learning Algorithms

Decision Tree. The commonly used DTs are party and rpart. In this work, after reading and partitioning, we have used a party package to implement conditional inference trees, which embed tree-structured regression models. We have also used ctree() from the package party to demonstrate the time series classification with the original data. Next, classification and regression trees were generated through the rpart package. It helps us explore the students' interest among those four fields.

Support Vector Machine (SVM). Firstly, the saved data from the RStudio are read and then partitioned using createDataPartition() function provided by the caret package. The model is trained using trainControl() and SVM linear methods. The result of the trained method was shown as "SVM with Linear Kernel". This result meant our data is linearly separable. In this SVM model, a linear kernel is used to predict the good value of a student's career interest among four fields.

Multinomial Logistic Regression. MLR is used to predict student interest among four fields (WDF, ADF, GDF, and DSF) without ordering. At first, we have read the dataset from RStudio. We had four levels for the response variable, so we have used one (1) as a reference level in the relevel() function, which means WDF. We have then used the multinom() function from the nnet package for the training dataset. We have found that the model is producing some errors, which are finally converged. Then, by predicting on the test dataset, we found the probability of students' career interest in four fields.

4.6 Predicted Results' Evaluation

Predicted results were evaluated using a confusion matrix (CM). Each ML classifier was evaluated first using different parameters such as accuracy, precision, recall, and $F1$-score of CM. Then, comparison results among ML classifiers were evaluated with the same parameters.

4.7 Comparison

In this research, the comparison is used to justify that the selected features, and the applied algorithm is worthy of further research. The comparison is made among different statistical methods. The comparison is also made among different ML classifiers. Finally, the comparison is made between statistical and ML results.

5 Data Analysis

Data analysis is done using descriptive statistics (DS). Among many factors in our study, here, we have discussed only some essential factors. Findings of the study show that 36.4% of students want to be a web developer, 24.1% of students want to be an Android developer, 18.6% of students want to become a graphic designer, and 20.9% of students want to become a data analyst in his career as it is one of the demanding sectors in the current world.

Out of 250 students, 44% (115) of students are male and 54% (135) are female. About 74.8% (187) of students have little knowledge in WDF, and 63 (25.2%) have an adequate concept. About 44.8% (112) of students specify that they have little knowledge about algorithms or data structures. The rest of them have no idea about algorithms or data structures. Only 13.2% (33) of students have little knowledge in GDF, and 26.8% (67) of students have DSF ideas.

6 Results and Discussions

6.1 Statistical Results

Bivariate analysis of Chi-squared test of independence. The test of independence between WDF, ADF, GDF, and DSF and background characteristics of respondents are performed in this section. Moreover, it helps us identify those independent variables, which have a significant effect on WDF, ADF, GDF, and DSF. The dependent variable is categorized into two groups, yes or no. Here, independent variables—first impression about WDF, earlier skills in WDF, previous knowledge in WDF, career as a web developer—are found to be highly significant with dependent variable WDF.

For ADF, the variables—feeling using any application, familiar with Android Studio, interested in coding, interest in algorithms or data structures, a career as an Android developer—are found highly significant with ADF. In the case of GDF, the variables—career as a graphic designer, reading most interested skills and technologies—are found to be significant at 1% level related to the dependent variable GDF. Finally, for DSF, the variables, first impression about DSF, a career as a data scientist, earlier skills in DSF, knowledge about mathematic, statistics, and database are found to be significant.

Binary Logistic Regression. Here, explanatory variables were considered as it found effective on bivariate analysis. In LR, we usually report the odds ratio along with regression coefficients. However, as per the objectives, this research reports only the classification table and model summary of BLR analysis. Table 2 shows that WDF is the demanding sectors in the current world. The classification in Table 3 indicates that this subject has more demand than WDF, whereas the classification in Table 4 indicates that this subject has more demand than WDF and ADF.

Finally, as per Table 5, 13% of students want to be data scientists. It indicates that this subject has less demand than other subjects. The reasons of less demands are (1) respondent are beginner level and (2) DSF subject is offered to the last year students.

The model summary Table 6 comprises the Cox and Snell R^2 and Nagelkerke's R^2 values, which compute the explained variation. Nagelkerke's R^2 is the adjustment of Cox and Snell R^2, the latter of which cannot achieve a value of 1. Table 6 also includes the pseudo R^2; the -2 log likelihood is how well the model explains variations in the outcome of interest. The explained variation in the outcome variable based on our model ranges from 67.0 to 90.7% for WDF, 54.4 to 74.6% for ADF, 60.5 to 84.5% for GDF, and 42.3 to 78.0% for DSF subject to whether you reference the Cox and Snell R^2 or Nagelkerke's R^2 methods.

Table 2 Classification of WDF

	Observed		Predicted		
			web developer		Percentage correct
			No	Yes	
Step 0	Web developer	No	0	99	0.0
		Yes	0	151	100.0
	Overall percentage				60.4

Table 3 Classification of ADF

	Observed		Predicted		
			Android developer		Percentage correct
			No	Yes	
Step 0	Android developer	No	0	90	0.0
		Yes	0	160	100.0
	Overall percentage				64.0

Table 4 Classification of GDF

	Observed		Predicted		
			Graphic designer		Percentage correct
			No	Yes	
Step 0	Graphic designer	No	0	81	0.0
		Yes	0	169	100.0
	Overall percentage				67.6

Table 5 Classification of DSF

	Observed	Predicted			
		data scientist			Percentage correct
		No		Yes	
Step 0	Data scientist	No	217	0	100.0
		Yes	33	0	0.0
	Overall percentage				86.8

Table 6 Model "Summary" in logistic regression

Dependent variable	− 2 Log likelihood	Cox and Snell R^2	Nagelkerke's R^2
WDF	58.398[a]	0.670	0.907
ADF	130.154[a]	0.544	0.746
GDF	82.615[a]	0.605	0.845
DSF	57.684[a]	0.423	0.780

[a]Estimation terminated at iteration 20 because maximum iterations have been reached. The final solution cannot be found

6.2 *Machine Learning Results*

Results of Decision Tree. In Fig. 2, we found prediction results for the testing dataset. Here, *x*-axis represents the students' ID and *y*-axis represents the predicted fields of interest. The figure shows that students are interested in fields 1 which is WDF, some in ADF, and so on. However, students are mostly interested in GDF, as per Fig. 2.

Results of SVM. Its result is saved in a svm_Linear variable. It is a linear model; therefore, it just tested at value "*C*" = 1. Here, the predict() method was used for predicting outcomes. Two arguments were passed through that method: the first

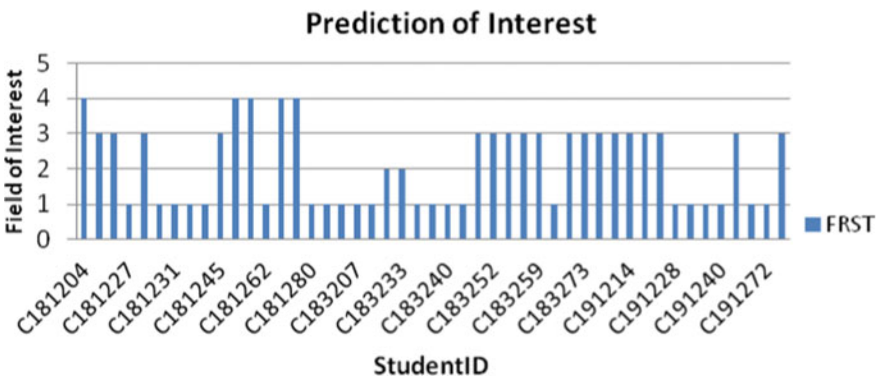


Fig. 2 Histogram for predicting field of interest with students' ID using decision tree

parameter for the trained model and the second parameter "newdata" holds the testing data frame. In Fig. 3, prediction results were found for the testing dataset. Here, x-axis represents the student ID, and the y-axis represents the predicted fields of interest. The result shows that students have diverse choices. However, considering the frequency, we may conclude that the students are highly interested in WDF.

Results of Multinomial Logistic Regression. MLR algorithm is used for finding significant variables, which was helpful in better prediction of interest. From the MLR model, we found the significance of variables. In Fig. 4, prediction results for the testing dataset are represented. Here, x-axis represents the student ID, and the y-axis represents four fields of interest. The data in the figure show that students are highly interested in field 1 which is WDF, and few students choose 3, GDF.

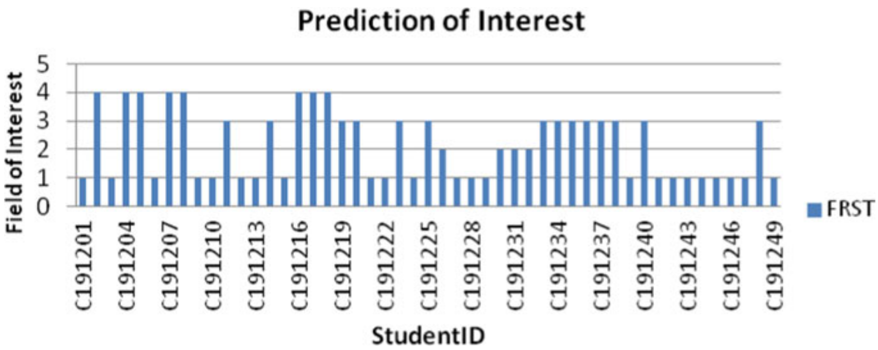


Fig. 3 Histogram for predicting field of interest with students' ID using SVM

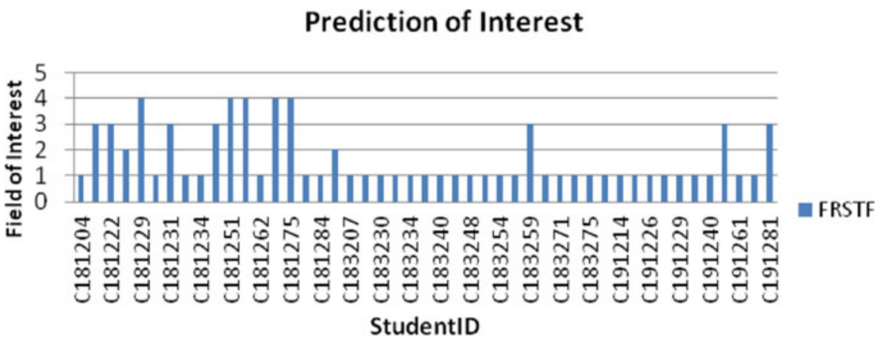


Fig. 4 Histogram for predicting field of interest with students' ID using logistic regression

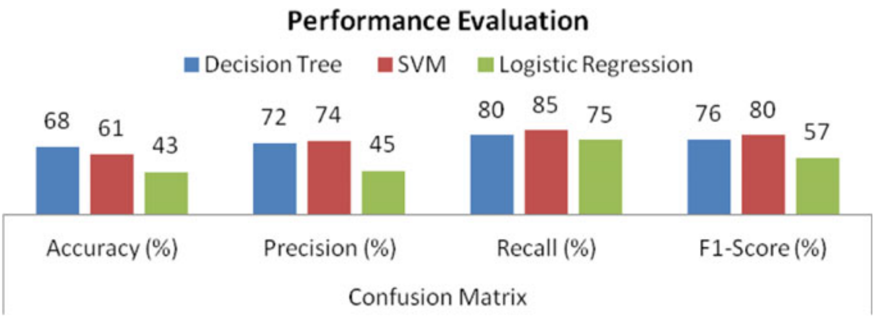


Fig. 5 Overall performance of the classifiers

6.3 Performance Comparison

Figure 5 indicates that on the scale of predictive accuracy, DT is outperforming other algorithms. The highest accuracy obtained is 68%. However, on the scale of overall performance, SVM is outperforming other algorithms. The highest precision, recall, and $F1$-score values are 74%, 85%, and 80%, respectively. It could be concluded that SVM and DT are interchangeably better, but BLR is lagging on the scales of all the parameters. In comparison with ML and SA, LR works better when applied in the SA. Therefore, we could conclude that BLR works better than MLR.

Overall, ML algorithms work better on the collected data than the SA. The comparison with the similar types of works, datasets, and algorithms indicates that this work, its related dataset, and algorithms show trustworthy results and are suitable for further research.

7 Conclusion and Future Work

Career choosing is a crucial factor that plays a vital role in a student's life. If a student lacks the right decision in career selection, their potential gets wasted. The idea of this work is to help beginner-level university students for finding opportunities to fill up the gap between their potentiality and the right career. The idea mentioned above is implemented by analyzing collected data from beginner-level university students with SA and ML techniques. These techniques predict a specific field of interest among WDF, ADF, GDF, and DSF. The analysis showed that students are highly interested in GDF and less interested in DSF. In future, we could collect the input data from other universities in Bangladesh besides IIUC to find the variation of response in a much better way.

References

1. Roy, K.S., Roopkanth, K., Teja, V.U., Bhavana, V., Priyanka, J.: Student career prediction using advanced machine learning techniques. Int. J. Eng. Technol. **7**(2.20), 26–29 (2018)
2. Gorad, N., Zalte, I., Nandi, A., Nayak, D.: Career counselling using data mining. Int. J. Eng. Sci. Comput. (IJESC) **7**(4), 10271–10274 (2017)
3. Ansari, G.A.: Career guidance through multilevel expert system using data mining technique. Int. J. Inf. Technol. Comput. Sci. **9**(8), 22–29 (2017)
4. Mishra, T., Kumar, D., Gupta, S.: Students' employability prediction model through data mining. Int. J. Appl. Eng. Res. **11**(4), 2275–2282 (2016)
5. Tarekegn, G.B., Sreenivasarao, V.: Application of data mining techniques to predict students placement in to Departments. Int. J. Res. Stud. Comput. Sci. Eng. (IJRSCSE) **3**(2), 10–14 (2016)
6. Nie, M., Yang, L., Sun, J., Su, H., Xia, H., Lian, D., Yan, K.: Advanced forecasting of career choices for college students based on campus big data. Front. Comp. Sci. **12**(3), 494–503 (2018)
7. Choe, N.H., Borrego, M.: Master's and doctoral engineering students' interest in industry, academia, and government careers. J. Eng. Educ. **109**(2), 325–346 (2020)
8. Lakin, J.M., Wittig, A.H., Davis, E.W., Davis, V.A.: Am I an engineer yet? Perceptions of engineering and identity among first year students. Eur. J. Eng. Educ. **45**(2), 214–231 (2020)
9. Al-Dossari, H., Nughaymish, F.A., Al-Qahtani, Z., Alkahlifah, M., Alqahtani, A.: A machine learning approach to career path choice for information technology graduates. Eng. Technol. Appl. Sci. Res. **10**(6), 6589–6596 (2020)
10. Laguador, J.M., Chavez-Prinsipe, N.H., De Castro, E.L.: Employability skill development needs of engineering students and employers' feedback on their internship performance. Univ. J. Educ. Res. **8**(7), 3097–3108 (2020)
11. Park, T., Kim, C.: Predicting the variables that determine university (re-) entrance as a career development using support vector machines with recursive feature elimination: the case of south korea. Sustainability **12**(18), 7365 (2020)
12. Fuchs, T.T., Scott, S.A., Sonnert, G., Sadler, P.M.: Predictors of high school students' interest in science and mathematics teaching careers. In: 2020 Annual Meeting, San Francisco, USA, pp. 1–12 (2020)
13. Miller, K.A., Sonnert, G., Sadler, P.M.: The influence of student enrollment in pre-college engineering courses on their interest in engineering careers. J. Pre-College Eng. Educ. Res. (J-PEER) **10**(1), 7 (2020)
14. Elmasry, M.A., Abdallah, A., Abran, A., Mohammed, A.K., Ogunbanwo, A., Okesola, J., et al.: Rationalizing the choice of academic major for undergraduates using classification algorithms. J. Theor. Appl. Inf. Technol. **99**(6), 1244–1256 (2021)
15. Hoff, K.A., Einarsdóttir, S., Chu, C., Briley, D.A., Rounds, J.: Personality changes predict early career outcomes: discovery and replication in 12-year longitudinal studies. Psychol. Sci. **32**(1), 64–79 (2021)
16. VidyaShreeram, N., Muthukumaravel, D.A.: Student career prediction using machine learning approaches. In: International Conference on Computing, Communication and Control (I3CAC-2021), pp. 1–8 (2021)
17. Hair Jr., J.F., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis: A Global Perspective, 7th edn. Pearson Education, Upper Saddle River (2010)

Factors Responsible for Fluctuations in Road Accidents: A Statistical Investigation of Road Accidents in Bangladesh

Shakera Khandekar, Sharna Das, and Sajjad Waheed

Abstract Road accidents have become an unwanted phenomenon in our daily lives. This paper aims to give a clear concept of the huge number of road accidents in the districts of Bangladesh. It also provides a detailed summary of the total accidents finding how different features are contributing to the increment of accidents and pointing out the accident-prone regions so that the losses can be minimized in accidents with limited resources. This paper presents various charts and graphs that show variation in the number of accidents based on different factors, victims, injured-to-dead ratio, locations of accidents in terms of number. The paper analyzes data on road accidents in Bangladesh. The experimental results show that the number of accidents is related to date, time, locations, and vehicle types. The results also reveal that 66.4% are injured and 33.6% are dead. It is a notable turning point over the past decades for this country. The result of this analysis exhibits different features that are important to reduce road accidents and to manage the transportation system of Bangladesh including pedestrians to impact public health and the economy of the country.

Keywords Road accidents · Data analysis · Bangladesh · Casualties

1 Introduction

This paper provides a unique dataset consisting of a variety of factors regarding road accidents in Bangladesh. This paper highlights the recent scenarios of road traffic

Supported by MBSTU

S. Khandekar (✉) · S. Das · S. Waheed
Mawlana Bhashani Science and Technology University, Santosh-1902, Tangail, Bangladesh
e-mail: it19617@mbstu.ac.bd

S. Das
e-mail: it19616@mbstu.ac.bd

S. Waheed
e-mail: sajad302@yahoo.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_6

Table 1 Involvement of heavy vehicles in fatal and pedestrian accidents

Variation of accidents	Bus or minibus (%)	Truck (%)
Heavy vehicles	33	27
Fatal accidents	35	29
Pedestrian accidents	38	30

accidents (RTAs) and the causes of the increasing number of accidents and their fatalities. An analysis is conducted using a simple data mining process to obtain data and illustrate the variety of road accidents with various factors. This paper also discusses possible reasons for the increasing number of accidents; based on this study, some recommendations are suitable to resolve the issues occurring with accidents. A massive number of death and injury caused by road accidents exposes the global crisis of road security. According to the WHO, the total population is decreased by 1.35 million and 20–50 million more tolerant non-fatal injuries, with many incurring a disability result of their injury every year. It shows that road traffic accidents cost most countries 3% of their gross domestic products. It estimates that 93% of the world's fatalities on the roads occur in low- and middle-income countries although these are the countries that have approximately 60% of the world's vehicles. According to the source, the victims of road accidents are primarily children and young adults aged between 5 and 29 [1]. From 2009 to July 2016, the rate of injury was 43.8% and the rate of the dead is 56.2%. Table 1 gives a summary of this data. As a developing country, road accidents have become a crucial concern in Bangladesh. In Bangladesh, road accident is identified as a major cause of death. Road and traffic accidents are unpredictable and dubious. Thus, it requires a plethora of information about the factors affecting them to analyze. This paper provides the information as thoroughly as possible based on the obtained dataset.

2 Literature Review

Based on a statistical analysis of road accidents from 2014 to 2016 in Bangladesh, Hossen (2019) found that road accident casualties have increased gradually under the calculation of excluding motorcycles compared to including them [2]. However, in another research paper, Kaur and Kaur (2017) compared state highways and ordinary district roads and types of accidents for the predictions that depicted relationships between the frequency and types of accidents [3]. Solaiman et al. (2013) conducted a study where a set of procedures for extracting information of interest for road safety evaluation were discussed. This started with data collection concerning road accidents. This approach processed spatial data to provide end-users with a reliable tool to evaluate what roads have the highest accident density, the highest danger rate and any other statistical indicator. All data could be extracted from the attributes, and the proposed tool was said to be the first software in Bangladesh, where all these methods and facilities were integrated [4].

Table 2 Percentage of casualties from the year 2009 to 2016

Year	Percentage of death	Percentage of injuries
2009	52.4	47.6
2010	59.5	40.5
2011	60.8	39.2
2012	54.3	45.7
2013	58.4	41.6
2014	57.4	42.6
2015	54.8	45.2
2016	52.5	47.5

Table 1, which was found in a research work conducted by Anjuman et al. in 2007, is studied on road accident variations [5]. This study indicated the cause of death due to road traffic accidents. The share of the number of buses and trucks was nearly 70% and for pedestrians about 72%. Alam et al. (2011) showed the safety situation in Bangladesh that had been presented in terms of the number of accidents and fatalities occurring each year. Subsequently, accidents and fatality rates in terms of unit registered motor vehicles had been evaluated as well [6]. Hence, Kumar and Toshniwal (2016) proposed a framework to analyze road accident time-series data that accumulated data from 39 districts and used a time-series merging algorithm and revealing that road accident trends were going to increase in certain clusters and those districts should be the prime concern to take preventive measures for mitigating road accidents [7]. On the contrary, Hossain and Faruque (2019) disclosed the trends of road accidents and their causalities, accident rates, and vehicle rates showing a clear picture of the different infected districts by road accidents. They also elaborated on the age of the affected people, type of accidents, and different clashes and measured the highest percentage of road traffic accidents (RTAs) [8]. Ahsan et al. (2011) mainly highlighted the general characteristics of fatal car accidents and attempted to establish the most common types of fatal accidents and causal factors [9].

Table 2 indicates the casualties recorded from the period 2009 to 2016 containing the number of accidents, deaths, and injuries in Bangladesh that are shown as percentages [10].

3 Data and Methodology

3.1 Data Source

These cases of studies gather data from various newspapers. From the aspect of Bangladesh, The Daily Star [1] holds the reputation of publishing valid information. Thus, this study retrieves necessary data from The Daily Star. This dataset intends to be a nationally representative sample from the annual accident reports in Bangladesh.

The fundamental dataset for the study contains traffic accident records from July in the year 2017 to June of the year 2018. The dataset includes the columns, namely years, months, days, hours, parts of the day or night times, district names, neighborhood names, street names, types of injuries, the number of deaths, reported victims, the vehicle involved, and vehicle types. For better understanding, the paper has attributed Morning time between 4.00 am and 12.00 am, Noon was from 12.01 pm to 02.59 pm, Afternoon was from 03.00 pm to 05.59 pm, Evening started at 6.00 pm and finished at 7.59 pm, and lastly, Nighttime was between 08.00 pm and 3.59 am. For enhancement of our dataset, we also included data for the years 2019 and 2020. In this paper, we also included data on the casualties as a result of road accidents in Bangladesh from 2019 to 2020. Our primary statistical analysis is done based on our fundamental dataset collected by authors, and a comparative analysis is shown with the casualties of accidents in recent years.

3.2 Methodology

Tools of Analysis Accident data are being analyzed years after years for the amelioration of traffic management and public safety. For this purpose, many have used various languages, methods, tools, and software. Data analysts use Python, Java, C++, R, Scala, MATLAB, Julia, TensorFlow, and SQL for the data mining process. All of these are well known. Many scientists are using these tools for analyzing data efficiently. Researchers use Python, R, MATLAB, TensorFlow, Orange in data-related works nowadays. As for software or tools, analysts use Anaconda, R, Orange, XNIME, XMiner, RapidMiner, WEKA. Python is an exceptionally popular and widely used language in the data science field. In the growing world of data, Python is booming for its efficiency and easy-going nature as a language.

As the dataset consists of both numerical and categorical values, the computing language used here is Python, a high-level but flexible data analysis language. The paper finds out necessary information from sources to visualize factors that fluctuate accident outcomes and discover trends and anomalies using the most efficient and simple analyzing technique and tools. On account of these, an inductive approach is appropriate for the proposed study. An inductive approach allows researchers to find the relevant information inherent in raw data. This approach consists of three parts: (i) gather information, (ii) mark the pattern, and (iii) develop a theory.

- This paper uses a dataset that is created by gathering information of road accidents in details day by day around the country from The Daily Star newspaper. Then, dataset is mined using various comprehensive libraries (Pandas, NumPy) of Python to generate a well-informed dataset.
- This study uses Python as a language and Anaconda Jupyter notebook as a working tool for the proposed analysis for various patterns of factors responsible for fluctuation in numbers of the accidents which are demonstrated with some graphs and charts.

- Lastly, through comparisons and discussions, the factors responsible for changes in casualties during accidents are justified and recommendations are provided.

This study mainly focuses on time, vehicle type, and location for road accidents and data analysis. From the newspaper, we have day, hour, month, weekday, part of the day as whole time information. A Python library term, matplotlib was used for illustrating the information about accidents. To illustrate figures, the columns are converted into integers and the "DateTime" function is used for finding the resolution of when and at what times accidents took place and in which parts of the day accident occurred in a substantial figure. From the main dataset, we have district names, neighborhood names, street names that represent the location of an accident, but all this information is an object type. Hence, two new columns named longitude and latitude were added by measuring the location using Google Map. The longitude and latitude of each car accident visualize the road location and its condition. The best way to analyze special data is by using maps. Folium is a Python library that helps create several types of Leaflet maps. We can easily generate a map of Bangladesh, creating a Folium map object.

4 Results and Discussion

4.1 Analysis of Accident Trends from 2017 to 2018

It is insisted that this paper contains the illustration of variety in the number of accidents based on the factors of newly collected accident data. As results, various graphs are shown based on the features of time, vehicle, and place.

Trends of the number of injuries and death over months: About 66.4% of total victims are injured and 33.6% are dead. This figure depicts that the death rate is less than the injured, but 33.6% is not an ignoring matter. Also, the injured rate is alarming and detrimental to the day-to-day life of the population (see Fig. 1).

Trends of the number of accidents over a week: The distribution of accidents occurred by day of the week and shows a peak on Saturday (beginning of the week in Bangladesh) which holds the highest number of accidents and Monday holds the lowest number. This graph depicts that on average, 79 accidents occurred per week throughout the year. It can be included as a cause that Saturday is the first day of the week and also the first working day of the week (see Fig. 2).

Trends of the number of accidents over hours: The distribution of accidents occurred by the hour of the day and shows a peak at 9 am and 8 pm, and again, the highest number of accidents occurs in 8–11 in the morning and in 7–11 in the night. Each column in the bar chart represents one hour. The number of accidents during the late night and early morning is lower than during other parts of the day, but there are a significant number of accidents occurring almost at all hours (see Fig. 3).

Trends of number of accidents over months: The distribution of accidents occurred by months which shows a peak in June and the lowest in February. The

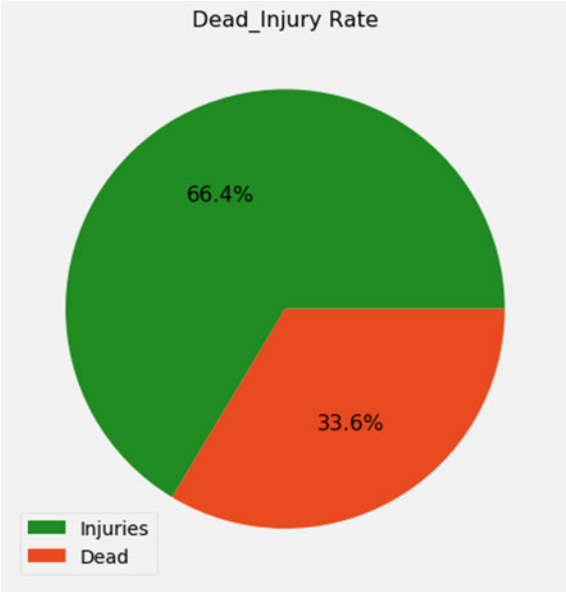


Fig. 1 Percentage of death and injury

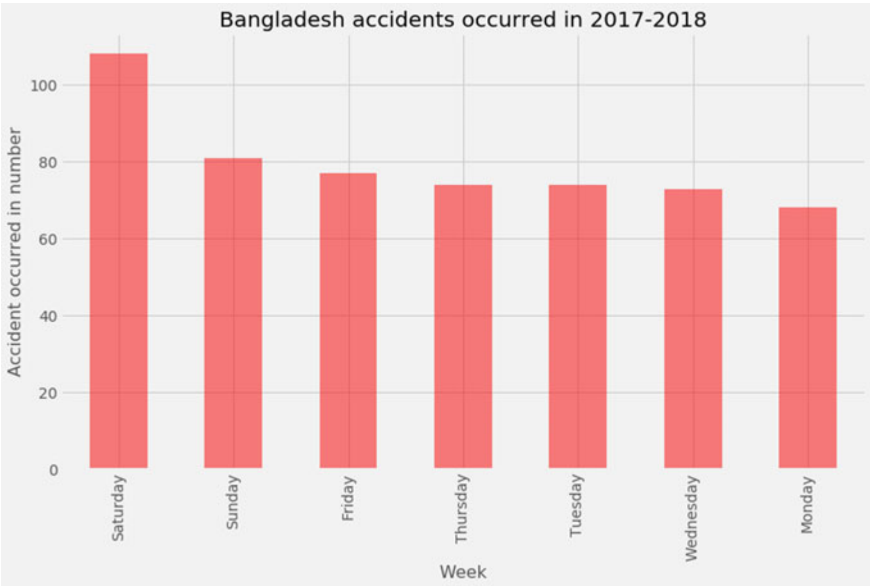


Fig. 2 Accidents per week

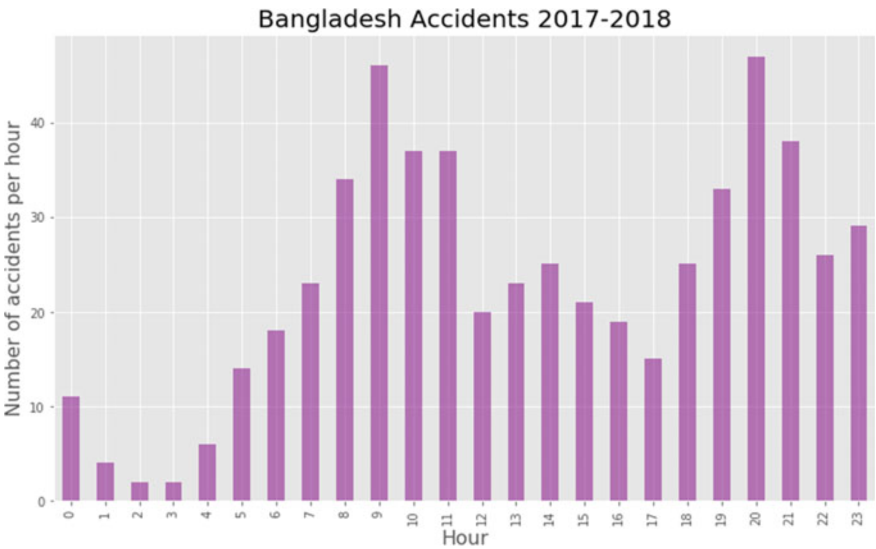


Fig. 3 Accidents per hour

total number of accidents in February is 29 and 61 in June. In the comparison of the last six months of 2017, more accidents occurred in the first six months of 2018. A maximum number of accidents occurred in June and a minimum number of accidents occurred in February. The potential cause of the increased number of accidents is the festival of Eid-Ul-Fitr (Table 3). A large number of working class people are garment workers, whereas another portion works in the divisional cities. They work in different garment factories and other organizations away from home. So, during Eid holidays, they travel back to their home districts. As Eid approaches, chaos breaks out on highways and roads and frequent movement of traffic causes an increase in the number of accidents (see Fig. 4).

Table 3 records casualties of thirteen days around Eid-Ul-Fitr in June containing date of accidents, number of accidents, deaths, and injuries that are shown as percentages [1].

Trends of accidents and vehicles: The distribution of accidents in terms of vehicle type is shown in Table 4. There are 35 ways in which vehicles are involved in different road accidents. It is shown that most accidents are caused by the bus (see Fig. 5).

The percentage of how many times one category of vehicles is involved in accidents. The types of vehicles involved in accidents are six-wheeler (47%), four-wheeler (5.6%), three-wheeler (22.7%), two-wheeler (18.3%), and other kinds of vehicles (6.4%). Also, from the graph, it is clear that the six-wheeler (47%) vehicles cause more accidents followed by three-wheeler vehicles (22.7%) (see Fig. 6).

Most accidents involve two vehicles. About 265 accidents have occurred involving two vehicles according to The Daily Star. Most of these are caused by a head-on collision. Also, it is quite interesting that the number of single-vehicle accidents

Table 3 Accidents in 13 days around Eid-Ul-Fitr

Month	Date	Number of accidents	Death	Injured
June	19/06/2018	12	20	92
June	20/06/2018	06	06	20
June	21/06/2018	12	16	18
June	22/06/2018	11	14	49
June	23/06/2018	15	18	35
June	24/06/2018	20	41	85
June	25 to 28/06/2018	60	80	128
June	29/06/2018	23	34	124
June	30/06/2018	21	23	78
July	01/07/2018	25	22	229
Total		205	274	848

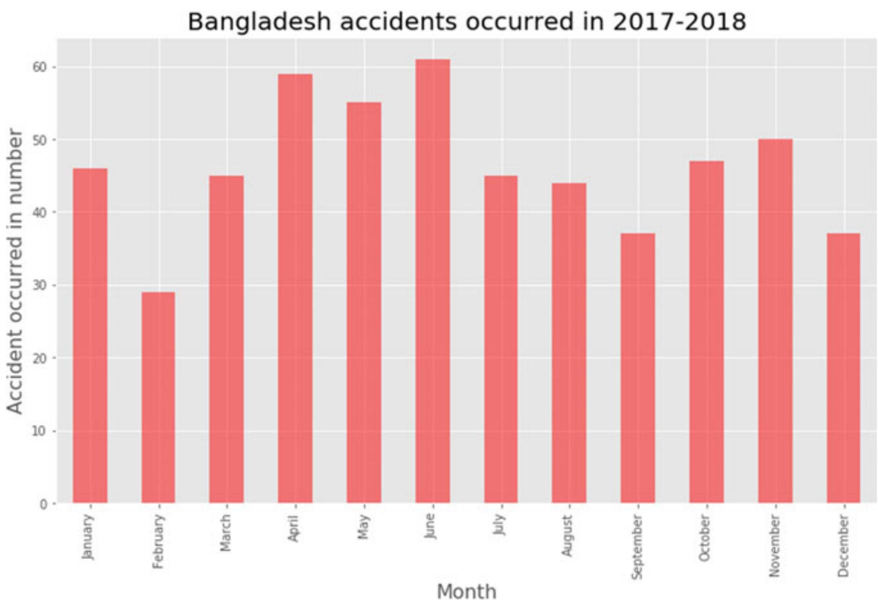


Fig. 4 Accidents per month

Table 4 Different categories of vehicle

Vehicle categories	Vehicles' name
Six-wheeler	Bus, truck, covered van
Four-wheeler	Private car, cargo van, hauler, ambulance
Three-wheeler	Auto rickshaw, CNG, van, rickshaw
Two-wheeler	Cycle, motorcycle

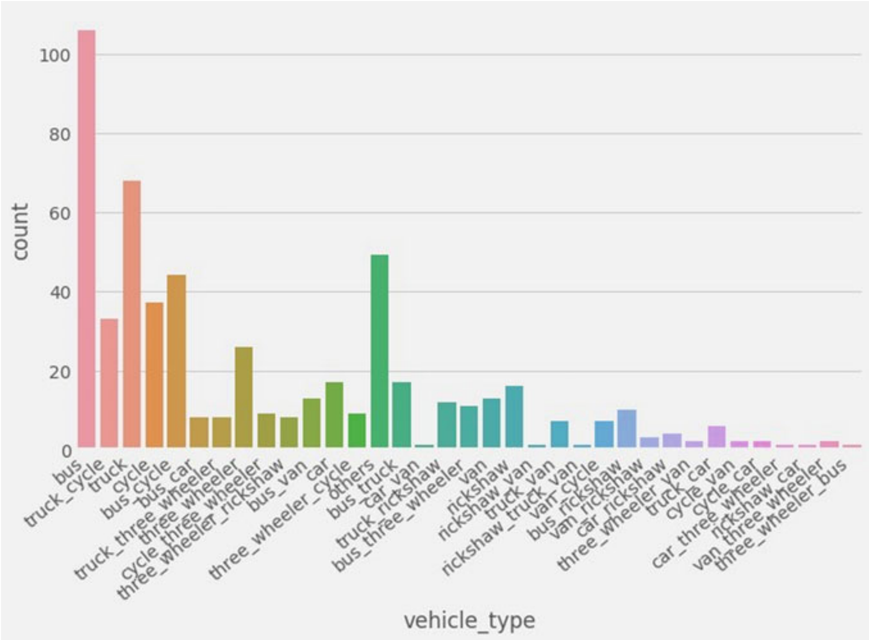


Fig. 5 Amount of different types of vehicles involved in accidents

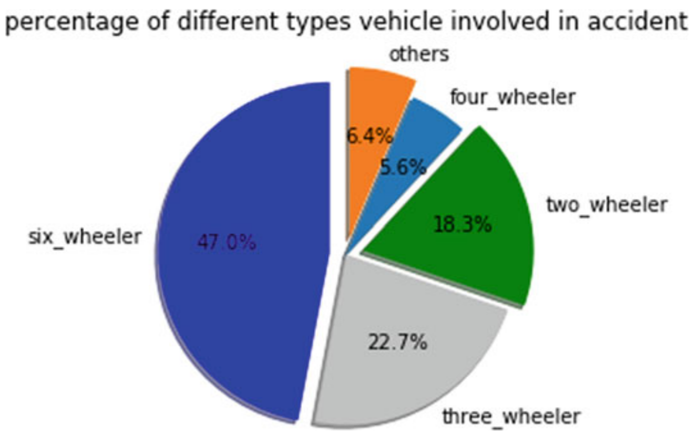


Fig. 6 Percentage of involvement of different categories of vehicles in accidents

is not less either being above 250. Most accidents were that it was quite normally involved with one or two vehicles. Accidents involving three vehicles could also be seen while it was quite rare to observe accidents that involved four or more vehicles. But, our analysis had shown that it was reported to have four vehicles involved in an accident in January 2018 (see Fig. 7).

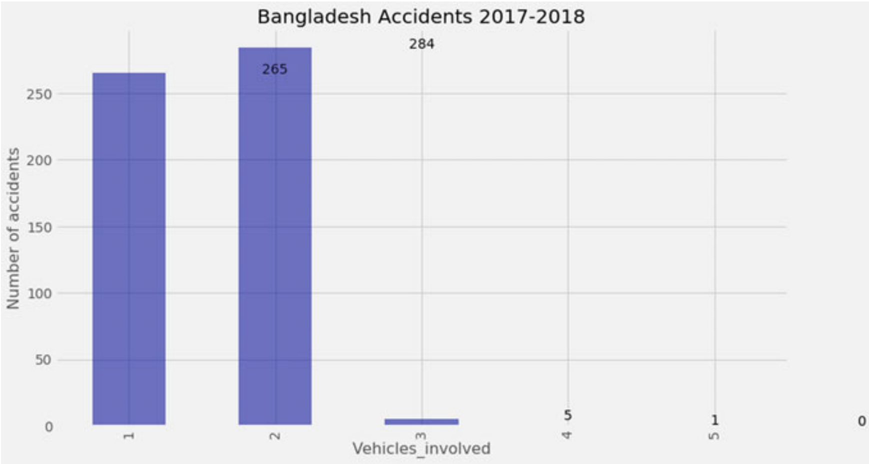


Fig. 7 Number of vehicles involved in accidents

Roads are inconveniently narrow in this country. So, collisions between two vehicles are a very common sight here due to small roads and defective traffic rules and huge population. But, the number of accidents involving one vehicle is also on the rise because of the poor condition of the roads and streets. It is a surprise that falling in ditches and crashing with trees are the main reasons for accidents involving one vehicle. This analysis has found that accidents are maximum by colliding two vehicles and maximum accidents involved six-wheeler vehicles.

Trends of accidents and locations: An essential part of the accident data system was location analysis. It pinpoints the accident locations in the Google Map based on the latitude–longitude illustrated (see Fig. 8). From this figure, the only division has the highest rate of accidents such as Dhaka city, Rangpur city, Chattogram city. We use the Folium library for showing the country map. Folium is a powerful Python library that helps to create several types of Leaflet maps. Using maps instead of other forms of charting allows us to highlight trends, uncover patterns, and reveal realities not visible before when it comes to spatial data. It also helps to gain clarity about the data, more than just simplifying the data itself (see Fig. 8).

This study uses Folium and MarkerCluster for identifying the location of the accident area where most of the fatalities have occurred. Figure 9 shows the district where most of the accidents happen. This represents regional accidental insights from the data.

Table 5 provides the information about total victims and the total death of the top seven districts out of 64 districts in terms of the casualties. Dinajpur district holds most victims according to our resources, and the total victim is 264 and 82 are dead among them. Chittagong is the second most accident-prone district with 228 victims and most dead victims (95) occurred in accidents here. Including these, Mymensingh, Faridpur, Tangail, Gopalganj, and Pabna are the cities with the most casualties which are shown in the following table.

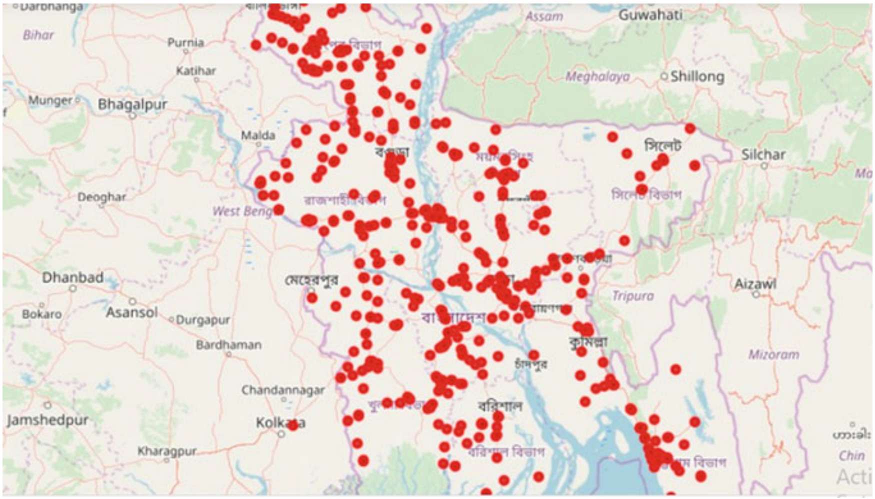


Fig. 8 Pinpointing locations of accidents

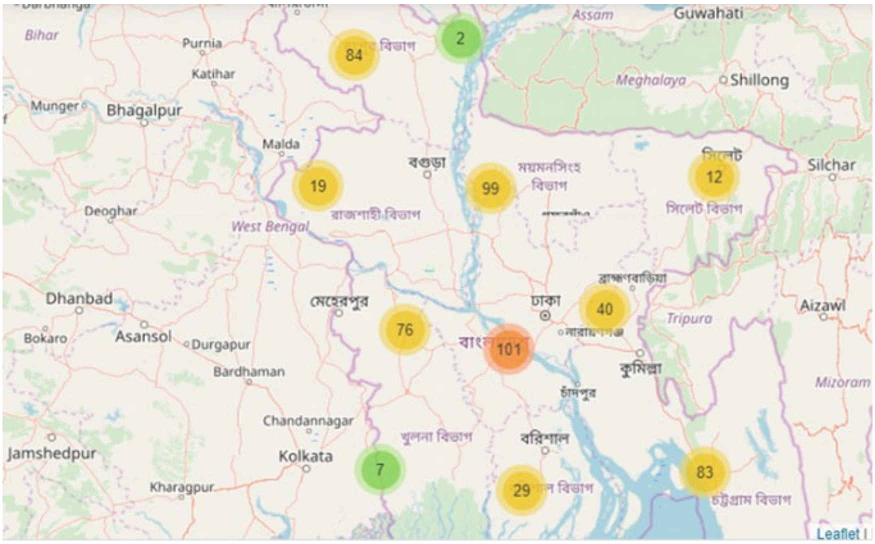


Fig. 9 Most fatal locations (districts) based on the number of victims

One of the most important parts of this paper is that it contains the location analysis of the accidents which is a big help to find out the risky highways and roads. Result analysis has shown that all the highways had taken a heavy toll on lives. It shows that the Dhaka–Mymensingh Highway, Dhaka–Chittagong Highway, Dhaka–Rangpur, and Dhaka–Gopalganj Highway have the maximum dead victim involved in accidents (see Fig. 8). Also, data analysis has shown the region of accidents by

Table 5 Top seven districts with the number of total victims and total

Name of the district	Total victims	Total death
Dinajpur	264	82
Chattogram	228	95
Mymensingh	185	55
Faridpur	108	31
Tangail	101	23
Gopalganj	97	35
Pabna	94	26

the cluster of the number of victims, and we have found out that the capital (Dhaka), the southwest of the country (i.e., Jessore, Gopalganj), and the north (i.e., Rangpur, Dinajpur) have highest number of victims as clusters who got involved in accidents (see Fig. 9).

4.2 Comparative Analyses of Accident Trends Between 2017–18 and 2019–20

From our analyses, we find that at the year 2017–18, the casualty rate is 33.6% death and 66.4% injured, whereas the casualty rate at the year 2019–20 is 43.4% death and 56.6% injured in our country. It clearly shows that the death rate has increased by 9.8% and the injury rate has decreased by 9.8% from the previous year which is a matter of great concern for us. Not only in the aspect of casualties but also at the year 2017–18 considering the involvement of vehicles, the six-wheelers are 47%, where at 2019, the involvement decreased to 43.07%. In the case of accident-prone areas, this paper exhibits that the accidents' rate is high in the highway regions increased by 6.78% in 2020 [1]. Also, the highest number of accidents occurs on the Chittagong–Dhaka–Mymensingh Highway in 2020, and at 2017–18, the regions were the Dhaka–Mymensingh Highway, Dhaka–Chittagong Highway, Dhaka–Rangpur, and Dhaka–Gopalganj Highway that have the maximum dead victim involved in accidents.

5 Conclusion and Recommendation

5.1 Conclusion

This paper illustrates a graphical representation of statistical analysis of road accident data which brings out the aspect of changes in the number of accidents due to various factors considering the reasons with the difference in the number of casualties at different locations. The casualty rate is different in each yearly period. Our study finds that in 2017–18, the rate is 66.4% injured and 33.6% dead, and in 2019–20, it is 43.4% death and 56.6% injured. Through our study, it shows clearly that the accident-prone regions or highways have not changed much to reduce the accidents rate to 5–10%. After comparing the data, we can say that road accidents could likely be reduced by 30–40% if the 111 recommendations of the National Road Safety Council were properly implemented, according to the Accident Research Institute (ARI). But unfortunately, it has not reduced much. So, some recommendations are given considering the present road conditions and traffic situations in Bangladesh. In terms of future work, we intend to deploy a substantial dataset with more features including weather, seasonal effect, speed of vehicles, and pinpointing exact locations and also provide real-time prediction and forecasting of accidents using data mining techniques.

5.2 Recommendation

Recommendation from the road traffic accident analysis, some recommendations are provided suitable for reducing accidents and their after-effects. Though the number of accidents should be decreased as Bangladesh is moving fast with the development of technology and economy around the world, in recent years, the occurrence of accidents is on a rampage. So, a few recommendations are:

1. Constructed separated lanes for various kinds of vehicles such as heavy vehicles, slow-moving vehicles, small vehicles, and so on.
2. Turning highways into four or six lanes and roads besides highways should be broadened also.
3. Building separate roads along highways for slow-moving vehicles, keeping the footpaths useable, and correcting the faults in road designs.
4. Placing speed detectors at a specific range to monitor the speed of vehicles and verify the traffic laws on highways and make alcohol detectors available to the traffic police.
5. Strict enforcement of the law, increasing awareness about traffic laws, banning mobile phones while going across roads and driving vehicles, and issuing fines if anyone is reluctant to use zebra crossings and footbridges where necessary.
6. Drivers of vehicles must be well trained and must have proper training and evaluation before issuing licenses to drivers. According to the Bangladesh Road Transport Authority (BRTA), among 3.3 million vehicles, only about 2 million of these vehicles have driving licenses; the remaining vehicles are driven by unlicensed drivers. So, the authority and political bodies must be strict on this account.

Acknowledgements This manuscript has not been published yet and is not even under consideration for publication elsewhere. The authors are grateful to those who have contributed to this research. And, the authors are explicitly grateful to Almighty God, parents, and respectable teachers.

References

1. https://www.thedailystar.net/
2. Hossen, M.S.: Analysis of road accidents in Bangladesh. Am. J. Transp. Logistics (AJTL) **2**, 9 (2019). (ISSN: 2637-6172)
3. Kaur, G., Kaur, H.: Prediction of the cause of accident and accident prone location on roads using data mining techniques. In: 8th ICCCNT 2017, IEEE—40222
4. Solaiman, K.M.A., Rahman, M.M., Shahriar, N.: Avra Bangladesh collection, analysis visualization of road accident data in Bangladesh. 978-1-4799-0400-6/13/$31.00 ©2013 IEEE
5. Anjuman, T., Hasanat-E-Rabbi, S., Siddiqui, C.K.A., Hoque, M.M.: Road traffic accident: a leading cause of the global burden of public health injuries and fatalities. In: International Conference on Mechanical Engineering 2007 (ICME2007)
6. Alam, M.S., Mahmud, S.M.S., Hoque, M.S.: Road accident trends in Bangladesh: a comprehensive study. In: 4th Annual Paper Meet and 1st Civil Engineering Congress (2011)
7. Kumar, S., Toshniwal, D.: A novel framework to analyze road accident time series data: J. Big Data **3**, 8 (2016). IEEE—40222. 8th ICCCNT 2017
8. Hossain, M.S., Faruque, M.O.: Road traffic accident scenario, pattern and forecasting in Bangladesh. J. Data Anal. Inf. Process. **7**, 29–45 (2019). http://www.scirp.org/journal/jdaip
9. Ahsan, H.M., Raihan, M.A., Rahman, M. (2011) A study on car involvement in road traffic accidents in Bangladesh. In: 4th Annual Paper Meet and 1st Civil Engineering Congress, Dhaka, Bangladesh, pp. 191–199 (2011)
10. http://data.gov.bd/dataset/road-accident-and-casualties-statistics-2009-2016

User Similarity Computation Strategy for Collaborative Filtering Using Word Sense Disambiguation Technique

Md. Samsuddoha, Dipto Biswas, and Md. Erfan

Abstract Collaborative filtering is a sophisticated recommendation system strategy that efficiently manipulates recommendations corresponding to preferences of users. Computation of user similarities is one of the features of the collaborative filtering strategy. This paper has presented a user similarity computation strategy for collaborative filtering using word sense disambiguation technique. The staple contributions of this research are utilizing associative database model for representing searched words of users uniquely and word sense disambiguation technique for removing ambiguities for understanding the exact preferences of users. In addition to, a self-tuning constant and a computation model have been also developed to compute similarities of users. The proposed approach has overcome all the prominent challenges against the recommendation system and performed very promisingly on two distinct datasets with about 93.40 and 96.10% accuracy. Moreover, the performance results and comparison demonstrate that the proposed user similarity computation model for collaborative filtering is comparatively optimal than other existing collaborative filtering approaches.

Keywords Recommendation system · Collaborative filtering · Associative database model · Word sense disambiguation

1 Introduction

Recommendation system (RS) is a knowledgeable strategy that efficiently manipulates recommendations of different preferred and demanded things for immense number of users [1]. RS always looks after the preferences of users depending on

Md. Samsuddoha (✉) · D. Biswas · Md. Erfan
Department of Computer Science and Engineering, University of Barishal, Barishal, Bangladesh
e-mail: sams.csebu@gmail.com

D. Biswas
e-mail: diptobiswas2016@gmail.com

Md. Erfan
e-mail: irfan.bucse@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_7

their demands. It treats like a predictor that can make appropriate assumptions on the interested and demanded items for the users to recommend [2]. There are plenty of RS based applications, though it is extensively noticeable in entertainment and electronic business world [3]. Among all of those recommendation system strategies, collaborative filtering is one of the most proficient [3].

Collaborative filtering (CF) is extra-ordinarily capable of preparing a comparison on different similar preferences of users [4]. Moreover, CF considers the relationships depending on the choices, demands, requirements, ratings, and preferences among users [5]. Generally, collaborative filtering faces various challenges such as finding ample homogeneous users called data sparsity [6], determining semantic relationships on similarities for proper assumptions, identifying profiles of users who are spurious [1]. Those challenges create discontentment along with diminishing the performance and accuracy of the CF based RS approaches [7]. Collaborative filtering approach is tremendously essential to measure the similar choices, interests, and preferences of the different users [8]. However, traditional CF approach is not highly effective when it is required to know the preferences of similar users on a newly launched items [6]. Even, it faces different types of cold star difficulties and fully dependent on user's provided ratings [9]. Many researches and implementations have endeavored to eliminate this problem during computing user similarities on their interests. Moreover, this section of the CF related to computing users similarities, requires more robust research [10]. This research is an effort to construct a user similarity computation strategy (USCS) for collaborative filtering using the word sense disambiguation technique. The proposed USCS considered both the searched activities of the users and provided ratings of the users. The proposed approach consists of several modules such as a supplementary associative database model (ADBM) [11] and the linguistics word sense disambiguation (WSD) [12] approach. The ADBM has been utilized to represent searched words uniquely achieved from user's activities. On the other hand, the WSD has eradicated ambiguities in the meanings of those words to understand the demands of various similar and dissimilar users properly. The ADBM and WSD have provided aids to calculate similarities by considering the searched activities of the users. In addition to, a self-tuning constant and a similarity computation model has been developed in the proposed USCS to calculate the similarities among users by considering their provided ratings. The staple contributions are

- Searched words for various items have been uniquely represented by an ADBM for analyzing all words individually with more preciseness to measure the similarities among users.
- Ambiguities among user's searched words have been eradicated by the WSD technique to understand the proper meanings of those searched words for various similar items by various similar users.
- A self-tuning constant (STC) and a similarity computation model (SCM) have been developed for the proposed USCS for CF for considering and manipulating ratings provided by the users.

All involved modules have been implemented properly to fulfill the mentioned contributions and approaches have been depicted with explanations in the proposed methodology section. The proposed USCS has been applied on two distinct datasets and achieved the 93.40–96.10% accuracy for all the datasets. The fundamental objective of this research is to propose a USCS for CF using the WSD technique that has comparatively increased its appropriateness in prediction and enhanced its reliability in terms of computing user similarities. The experimental result and comparison demonstrate that the proposed USCS is comparatively better, optimal, and acceptable. The rest of the paper has been organized as—Related Work in Sect. 2, Proposed Methodology in Sect. 3, Results and Evaluation in Sect. 4, and eventually Conclusion in Sect. 5.

2 Related Work

Recommendation systems refer to recommending various items by predicting and analyzing the interests, choices, and preferences of users [4]. Collaborative filtering is the prominent RS technique that has the capabilities of identifying user's choices with a great accuracy. Even, it has an excellent and robust predicting ability [5]. Many researchers had developed enormous number of CF techniques for computing user similarity. This section has represented some existing collaborative filtering implementations.

Zhao and Shang [7] eliminated scalability problem by implementing a CF approach. Their CF approach was used based and personalized. Similar choices, interests, and preferences for various items have been identified by this CF strategy. However, this CF was unable to predict user similarities by analyzing choices, interest, and preferences of users. Chen et al. [8] considered profiles of various users to develop a CF approach as a RS. Their CF also considered attributes and activities of various users to find out the similarities among users. The implemented CF was user-dependent and personalized. However, this approach faced sparsity complications extensively. Jin et al. [6] introduced item-based CF strategy as a RS. The implemented CF was highly efficient in terms of determining similar choices, and preferences for various available items for users. However, this CF was unable to predict the similar demands, choices, and preferences of various users. Even, this CF could not measure semantic relationships for proper assumptions. Deng et al. [9] considered vector representations of the features of the latent of items during developing a CF approach. The developed CF manipulated factorization strategies for various types of matrix of the vectors. This approach was item-based, personalized, and suffered sparsity complications. Avesain et al. [10] developed a CF approach depending on trusts of users. The implemented CF method was trust-oriented and ski-mountaineering. This CF had an availability of expressing facilities of trusts with its scores. However, this CF demanded manual labors and defining appropriate trust scores. Lathia et al. [13] also considered trusts of the users with a higher degree during developing a CF approach. The implemented CF approach also considered user's odd and even ratings separately. This CF also suffered sparsity complications and semantic relationships complexities during assumptions. O'Donovan and Smyth [14] implemented a CF technique that had considered two different types of ratings such as predicted and actual values of ratings and a difference of those two ratings. It had followed transitivity and asymmetric features and faced sparsity complexities as well. Above CF studies introduced enormous number of complications and challenges in the recommendation system area. Handling sparsity complications, determining semantic relationships on similarities for proper assumptions, identifying spurious profiles of users are the most prominent challenges that diminish accuracy of a CF technique apparently.

3 Proposed Methodology

This section has explicitly clarified and explained the proposed user similarity computation strategy for collaborative filtering. The proposed implementation has been developed by integrating several subsections such as pre-processing, designing an associative database model, applying the word sense disambiguation technique, defining a self-tuning constant, developing a similarity computation model. Prominent modules have been individually depicted in Fig. 1 with a little indication and a few information.

Figure 1 has demonstrated the flow of working among the integrated subsections for the proposed USCS. The execution has been commenced from right-top corner by inputting IP addresses of the systems, searched results of the users, and provided

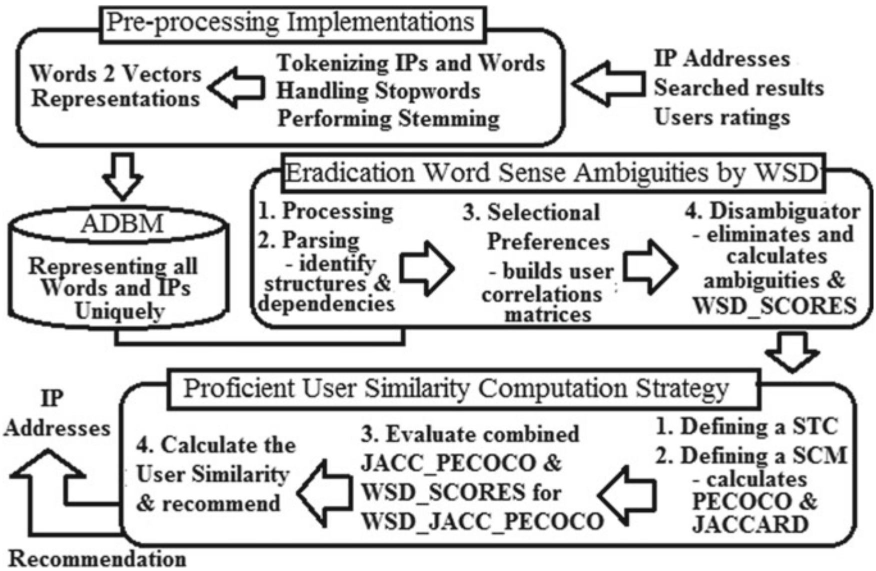


Fig. 1 Workflow of the proposed USCS for CF using WSD

ratings of the users. The proposed USCS has considered both the searched results and provided ratings of the users and computed similarities of the preferences from those two parameters. Searched results have been executed and analyzed by the ADBM and WSD techniques after pre-processing to calculate the user similarities depending on their demands. Moreover, provided ratings have been analyzed by a STC and eventually, those ratings have been executed by a SCM to compute the user similarities based on their preferences. With a view to understanding the proposed USCS for CF using WSD, all modules have been distinctly and elaborately described below.

3.1 Pre-processing

The pre-processing module has performed some basic operations before utilizing searched words of different users. IP addresses of the systems of the users, searched activities of the users, and provided ratings of the users have been accepted by the pre-processing module. The pre-processor has basically performed splitting IP addresses, tokenizing words, handling stop words, stemming, and transforming all required words into a word 2 vector representation. **Splitting IP addresses**: The pre-processing unit has split the IP addresses of the systems of the users and sent those to the word dictionary associated with the integrated ADBM. **Tokenizing words**: Searched sentences are generally large or short strings [11]. All have been tokenized into particular words for further representation in ADBM. **Handling stop words**: Searched sentences are full of linking and stop words and even, those words do not carry any special meanings [11]. Removing stop words technique has been applied here to handle all the unnecessary stop words. **Stemming**: Searched sentences contain gerunds, different types of past continuous, past perfect continuous sentences [11]. All those tokens contain same meanings but their forms of representation are not similar [12]. Since those tokens have almost same meaning, the stemming operation has been performed on them to convert them into a common parts of speech. This research has converted all the available valid tokens into a noun by the stemming operation. **Word 2 Vector representation**: It is really hard to identify a linguistic word directly by a computing system [11]. The integrated Word 2 Vector representation technique has assigned numerical values to all the tokens for easy identification. Those assigned values has provided additional advantages during representing those tokens into the integrated ADBM uniquely.

3.2 Unique Arrangement of Searched Words Through ADBM

Demands and requirements of users belong to searched words of users [5]. After pre-processing, all valuable words and IP addresses have been stored in a word dictionary. Then, all of those words have been represented uniquely with their corresponding IP

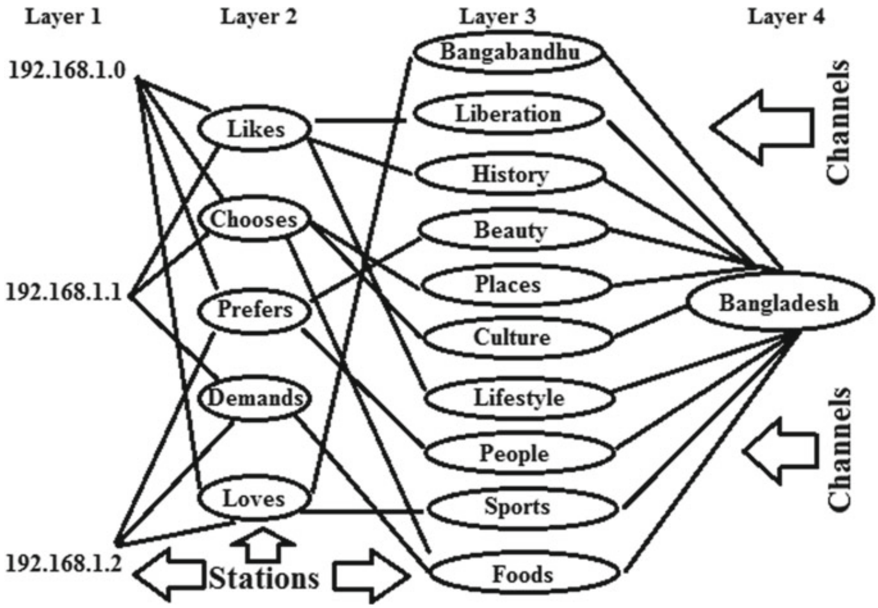


Fig. 2 Unique representation of searched words through ADBM

addresses through an ADBM. Representations have been depicted in Fig. 2 with an example. The example consists of 3 search histories of 3 users such as follows

User IP: **192.168.1.0**
History: liberation war of Bangladesh, Bangabandhu and liberation war of Bangladesh, history of Bangladesh, beauties of Bangladesh, historical places of Bangladesh, tradition and sports of Bangladesh

User IP: **192.168.1.1**
History: foods of Bangladesh, lifestyle of Bangladeshi, independence of Bangladesh, history of liberation war of Bangladesh, different types of people in Bangladesh, culture of Bangladesh, liberation war of Bangladesh, Bangabandhu

User IP: **192.168.1.2**
History: struggle of Bangladeshi people in 1971, life of Bangladeshi people during liberation war, role of Bangabandhu in liberation war, perception on Bangabandhu, Bangabandhu and Bangladesh, culture, sports, and tradition of Bangladesh, Bangabandhu, his role, and his significance.

Figure 2 has demonstrated the unique representation of searched words and IP addresses of 3 different users with some similar and distinct preferences. The integrated associative database model has two prominent data structures named stations and channels. **Stations** are normally principle or staple searched words of users represented as nodes. and **Channels** are generally links or paths that define associations and connections among all stations. In this ADBM, a set of predefined stations have

Table 1 Composition of the user correlation (UC) matrix

Users	User 1	User 2	User 3	.	.	User n
User 1	0	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$
User 2	$0 \le S_n < \infty$	0	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$
User 3	$0 \le S_n < \infty$	$0 \le S_n < \infty$	0	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$
.	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	0	$0 \le S_n < \infty$	$0 \le S_n < \infty$
.	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	0	$0 \le S_n < \infty$
User n	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	$0 \le S_n < \infty$	0

been utilized in the second layer depicted in Fig. 2. Those predefined stations can be or cannot be any searched words. Such type of unique representation of the searched words through ADBM has been extremely helpful to identify the similar users. Moreover, it has provided significant aid to eradicate ambiguities of the searched words by WSD approach.

3.3 Eradication of Word Sense Ambiguities Through WSD

A linguistic WSD technique has been utilized to eliminate ambiguities in searched words. To understand or identify the exact demands and preferences of the similar users, it is highly fruitful to reduce the ambiguities in meanings of their searched words [2]. The WSD strategy has offered two fold advantages such as identifying correlations among searched words, and reducing ambiguities among them. The integrated WSD in proposed USCS has 4 components such as a processing unit, a parsing unit, selectional preferences, and a disambiguator. The processing unit has split verbs, pronouns, nouns, adverbs, adjectives, and other parts of speeches. On the other hand, the parsing unit has imposed a well-defined grammatical structure on various parts of speeches to recognize the grammatical dependencies and structures. After that, the selectional preferences (SP) has been utilized in order to identify the correlations among all of those words. It has generally picked a word and explored relations among other words. The integrated SP has generated matrices for defining the correlations among all of the words for all the users individually. The composition of the matrix has been presented in Table 1, and the user correlation matrix has been defined as UC.

The above UC matrix has been further used for calculating WSD scores. Here, S_n is the number of similarities among various users on their preferences that belong to [0,

∞). Eventually, the disambiguator eradicates ambiguities with respect to their word meanings exactly. It also eliminated the redundant utilizations of similar type words from similar or dissimilar users. This research has eradicated the ambiguities of verbs, adjectives, and nouns considering nouns as base components by the disambiguator. The required equations are as follows.

$$\sum_{\mathrm{nc}\in\Gamma}^{\infty} P(\mathrm{nc}) = 1 \tag{1}$$

$$\sum_{\mathrm{nc}\in\Gamma}^{\infty} P(\mathrm{nc}|\mathrm{vc},\mathrm{gr}) = 1 \tag{2}$$

$$\sum_{\mathrm{nc}\in\Gamma}^{\infty} P(\mathrm{nc}|\mathrm{ac},\mathrm{gr}) = 1 \tag{3}$$

$$P(\mathrm{vc}|\mathrm{nc},\mathrm{gr}) = P(\mathrm{nc}|\mathrm{vc},\mathrm{gr}) * \frac{P(\mathrm{vc}|\mathrm{gr})}{P(\mathrm{nc}|\mathrm{gr})} \tag{4}$$

$$P(\mathrm{ac}|\mathrm{nc},\mathrm{gr}) = P(\mathrm{nc}|\mathrm{ac},\mathrm{gr}) * \frac{P(\mathrm{ac}|\mathrm{gr})}{P(\mathrm{nc}|\mathrm{gr})} \tag{5}$$

All of those equations have been utilized by the disambiguator to eliminate the ambiguities. Here, nc , gr, ac, and vc have defined noun-classes, grammatical-rules, adjective-classes, and verb-classes, respectively. Equations 1, 2, and 3 has defined proper classes of nouns, proper classes of verbs, and proper classes of adjectives, respectively. The summation of the probabilities of all classes have been considered 1. Eventually, Eqs. 4, and 5 has been utilized to calculate the probabilities of verbs, and adjectives by considering nouns as the base words. After pre-processing the WSD technique along with ADBM has computed the user similarities depending on their searching. The user similarities have been calculated by considering the following equation.

$$\mathrm{WSD\_SCORES} = \mathrm{Det(UC)} * P(\mathrm{ac}|\mathrm{nc},\mathrm{gr}) * P(\mathrm{vc}|\mathrm{nc},\mathrm{gr}) \tag{6}$$

The above equation has defined the WSD_SCORES that are numerical values. Those numerical WSD_SCORES have defined the obtained user similarities depending on their demands and considering the searching results of the users. Here, Det(UC) is the determinant of the user correlations matrix. $P(\mathrm{ac}|\mathrm{nc},\mathrm{gr})$, and $P(\mathrm{vc}|\mathrm{nc},\mathrm{gr})$ are the probabilities of the adjective words and verb-related words, respectively, achieved from Eqs. 4 and 5. Meanwhile, the STC and SCM have been applied to analyze and calculate the user similarities by considering the provided ratings of user described in the following next two subsections.

3.4 Applying of a Self-tuning Constant (STC)

Requirements, interests, demands, and preferences extensively vary among users [1]. The variation depends on current preferences and changes of preferences over time [2]. Some preferences stay for a long time and some preferences stay for short time [3]. Therefore, there must exist a Decay Rate (DR) which is inextricably dependent on user choices on specific thing. The DR provides aid to identify less important ratings and demands of users [4]. This identification is highly fruitful to ignore all the less important ratings [5]. This research has a special and crucial constant ψ has been utilized as the DR that has defined the variation of choices and preferences among different similar or dissimilar users.

$$Z = Z_R - Z_I \tag{7}$$

$$\psi = \frac{1}{Z} \tag{8}$$

Z_R is the time of recently provided rating by a user, and Z_I is the time of provided rating for the time I by that user. Higher DT value is not acceptable because it creates hindrances in prediction of similarities. The DT value ψ will be high if Z is low. Therefore, items that are old rated have been considered as less important item during assumption for computing similarities.

3.5 Developing the Similarity Computation Model

The SCM has been utilized to compute and identify users who have similar choices, demands, requirements, and preferences. This SCM has also considered those users as neighbors of each other's. The SCM has considered two types of users such as aimed users U_P and other users in the system except aimed users U_Q. Here $U_P \neq U_Q$ and even $U_Q \in U$ here, $U = \{U_1, U_2, U_3, \ldots, U_n\}$. There exists two efficient methods that have been integrated in the developed SCM. One is JACCARD module for similarity calculation and another is Pearson correlation coefficient (PECOCO). Firstly, The PECOCO has measured similarities among all the users numerically by considering their provided ratings on various items. Secondly, the JACCARD module has been applied to determine the similarities among all of those users by consider those different various items that are co-rated by the users. The equations for PECOCO and JACCARD have been mentioned as follows.

$$\forall = I_P \cap I_Q \tag{9}$$

$$\text{PECOCO}(P, Q) = \frac{\sum_{X\in\forall}^{\infty}(R_{P,X} - R'_{P,X}) * (R_{Q,X} - R'_{Q,X})}{\sqrt{\sum_{X\in\forall}^{\infty}(R_{P,X} - R'_{P,X})^2} * \sqrt{\sum_{X\in\forall}^{\infty}(R_{Q,X} - R'_{Q,X})^2}} \tag{10}$$

$$\text{JACCARD}(P, Q) = \frac{|P_i \cap Q_i|}{|P_i \cup Q_i|} \tag{11}$$

Here, $R_{P,X}$ are the provided ratings of U_P, and $R'_{P,X}$ are the average provided ratings of U_P to all the preferences of the user U_P. All ratings belong to 1–5. Similarly, $R_{Q,X}$, are recommendations to the U_Q for preferred items X, and $R'_{Q,X}$ are average recommendations to the U_Q for all preferred items X. As well as, the P_i are the number of preferred and demanded items for the U_P, and the Q_i are the number of preferred and demanded items for the U_Q, respectively. To achieve accurate and exact similarities, the results of PECOCO (P, Q) and JACCARD (P, Q) have been combined with a multiplication relation for the proposed PUSCS denoted by JACC_PECOCO.

$$\text{JACC\_PECOCO}(P, Q) = \text{PECOCO}(P, Q) * \text{JACCARD}(P, Q) \tag{12}$$

Normally, provided ratings to any items of the users define the preferences of the users to those items [1]. The obtained JACC_PECOCO values are numerical values that have defined the obtained user similarities depending on the provided ratings of the users. This obtaind JACC_PECOCO values and the WSD_SCORES have been further combined to achieve an appropriate user similarities from this proposed USCS.

3.6 Combination of WSD_SCORES and JACC_PECOCO

This section has integrated the WSD_SCORES and JACC_PECOCO to find out the appropriate similarities among the users. The WSD_SCORES and JACC_PECOCO have been basically integrated between the U_Q who are recommendation providers and the U_P who are aimed users. An arithmetic mean operation has been performed to get the comparatively appropriate user similarities from the integration of the WSD_SCORES and JACC_PECOCO.

$$\text{WSD\_JACC\_PECOCO} = \frac{\text{WSD\_SCORES} + \text{JACC\_PECOCO}(P, Q)}{2} \tag{13}$$

The combination of the WSD_SCORES and JACC_PECOCO has been defined by the WSD_JACC_PECOCO. This WSD_JACC_PECOCO is the final computation for the user similarities, and this calculation has provided comparatively better outcomes. The performance results of the WSD_JACC_PECOCO along with other sub-functions have been demonstrated in the result and discussion section elaborately.

4 Result and Discussion

4.1 Datasets

The proposed USCS for CF using WSD has been examined by using an own created movie-related dataset and a popular existing Movielens dataset developed by IMDB [1]. The own created movie-related dataset has been defined as DS1, and the existing Movielens dataset has been defined as DS2. Both datasets contain ratings and searched records related to movies. Here, both datasets consist of 1,000,209 and 100,000 ratings of 6040 and 600 users to 3952 and 300 movies with their searched results, respectively. All users have provided ratings at least 52 movies within a range 1–5.

4.2 Evaluation Metrics

The proficiency and appropriateness of the proposed PUSCS for CF using WSD have been measured by applying mean absolute error (Mean_Abs _Err). The Mean_Abs_Err has determined deviations of the assumed ratings and actual ratings of different similar users for different various items.

$$\text{Mean\_Abs\_Err} = \frac{\sum_{X=1}^{N_A} |R_{P,X} - P_{P,X}|}{N_A} \tag{14}$$

The above equation is for each user available in the system. N_A defines the number of movies. $R_{P,X}$ is the actual ratings of the users, and $P_{P,X}$ is the assumed ratings of the systems. Higher values of Mean_Abs_Err define lower proficiency and appropriateness of the recommendation system. On the contrary, small values of Mean_Abs_Err express and clarify that the recommendation system is highly and comparatively better, optimal, appropriate, and proficient.

4.3 Performance Results

The proposed USCS for CF using WSD has performed comparatively very well for both two datasets. The performance results of the proposed USCS have been demonstrated along with the results of other various sub-functions such as WSD_SCORES and JACC_PECOCO. The WSD_SCORES have been recorded into Table 2. and the JACC_PECOCO (P, Q) results have been achieved by analyzing the provided ratings of the users to any items recorded into Table 3.

The first column of Table 2 contains size or quantities of various neighbors. The other columns successively contain Det (UC), values of $P(\text{ac}|\text{nc, gr})$, and values

Table 2 Performance results of the WSD_SCORES for the proposed USCS

K-NB	Det (UC)		P(ac|nc, gr)		P(vc|nc, gr)		WSD_SCORES		Percentages	
	DS1	DS2	DS1	DS2	DS1	DS2	DS1	DS2	DS1 (%)	DS2 (%)
(60, 70)	15.02	14.99	0.200	0.200	0.320	0.303	0.959	0.910	**95.9**	**91.0**
(70, 80)	15.09	15.04	0.199	0.199	0.319	0.305	0.958	0.915	**95.8**	**91.5**
(80, 90)	15.14	15.10	0.198	0.199	0.319	0.311	0.961	0.934	**96.1**	**93.4**

Table 3 Performance results of the JACC_PECOCO (P, Q) for the proposed USCS

K-NB (P, Q)	PECOCO (P, Q)		JACCARD (P, Q)		JACC_PECOCO (P, Q)		Percentages	
	DS1	DS2	DS1	DS2	DS1	DS2	DS1 (%)	DS2 (%)
(60, 70)	2.053	2.036	0.457	0.438	0.939	0.890	**93.9**	**89.0**
(70, 80)	2.058	2.046	0.459	0.447	0.945	0.915	**94.5**	**91.5**
(80, 90)	2.065	2.055	0.465	0.455	0.961	0.934	**96.1**	**93.4**

of P(vc|nc, gr) for both datasets. The ultimate column represents the result of the WSD_SCORES in percentages about 84.2–93.4% for DS2. The WSD_SCORES values have provided the results of users similarities by analyzing the searched activities of the users and the provided ratings of the users have been analyzed by the STC and SCM to achieve the JACC_PECOCO (P, Q).

The first column of Table 3 contains size or quantities of various neighbors. The other columns successively contain PECOCO (P, Q), JACCARD (P, Q), and JACC_PECOCO (P, Q) results. The ultimate column represents the result of JACC_PECOCO in percentages for both DS1, and DS2. The percentages column represents the level of appropriateness of PUSCS for CF about 89.5–96.1% for DS1 and about 84.2–93.4% for DS2. This JACC_PECOCO values have provided the user similarities by considering the provided ratings of the users. Eventually, the WSD_SCORES and the JACC_PECOCO have been combined in order to get the appropriate user similarities. The combination has been defined as WSD_JACC_PECOCO and the results of the user similarities have been recorded in Table 4 by considering their involvements as weights.

The first column of Table 4 contains the weights of the WSD_SCORES, and the JACC_PECOCO as pairs. The other columns contain the results of the WSD _SCORES, the JACC_PECOCO, and the WSD_JACC_PECOCO, respectively, for both datasets. The last column contains the final user similarities with percentages for both datasets. The proposed USCS has provided user similarities about 96.10% for DS1 and 93.40% for DS2, respectively, achieved by the 100% involvement of the WSD_SCORES, and the JACC_PECOCO. This research has computed

Table 4 Combined results of the WSD_SCORES and the JACC_PECOCO

Weights	WSD_SCORES		JACC_PECOCO		WSD_JACC_PECOCO		Percentages	
	DS1	DS2	DS1	DS2	DS1	DS2	DS1 (%)	DS2 (%)
(0, 1)	0	0	0.961	0.934	0.4805	0.4670	48.05	46.70
(1, 0)	0.961	0.934	0	0	0.4805	0.4670	48.05	46.70
(1, 1)	**0.9610**	**0.9340**	**0.9610**	**0.9340**	**0.9610**	**0.9340**	**96.10**	**93.40**

Table 5 Comparative results of the proposed USCS and other CF approaches

K-NB	Proposed USCS		TJPCC [1]		TFS [15]		JMSD [16]		O'DT [14]		RUCF [17]	
	DS1	DS2	DS1	DS2	DS1	DS2	DS1	DS2	DS1	DS2	DS1	DS2
60	0.411	0.444	0.640	0.660	0.671	0.693	0.765	0.773	0.752	0.763	0.779	0.769
70	0.400	0.430	0.631	0.652	0.670	0.690	0.762	0.771	0.751	0.762	0.774	0.764
80	0.400	0.414	0.621	0.641	0.660	0.683	0.760	0.770	0.750	0.760	0.770	0.764
90	0.400	0.401	0.610	0.630	0.651	0.670	0.760	0.770	0.750	0.760	0.761	0.764

the WSD_JACC_PECOCO values as the final results of the user similarities, and it has been comparatively optimal. The proposed USCS has also been applied on other existing datasets and even on other own created datasets such restaurant rating datasets, traveling-group ratings datasets, and personality-review rating datasets, and so on. The proposed USCS has also applied on both low volume and high volume datasets and provided the level of appropriateness and proficiency almost about 93.40–96.10% for all datasets.

4.4 Discussion and Evaluations

Collaborative filtering involves and faces various challenges in terms of increasing its performance accuracy [1]. The proposed USCS for CF using WSD has eradicated various RS-related challenges such as finding ample homogeneous users called data sparsity, determining semantic relationships on similarities for proper assumptions. The proposed USCS has also considered a STC that creates great influence to avoid less important ratings. All of those involvements and features have increased the appropriateness of the proposed USCS. Table 5 has demonstrated the level of appropriateness and proficiency by representing the Mean_Abs_Err values of the proposed USCS.

Table 5 also has displayed a comparison of the proposed USCS with other existing popular CF techniques for both two datasets. The first column of Table 5 contains size or quantities of various neighbors. The second and third column holds the values of Mean_Abs_Err of the proposed USCS for DS1 and DS2, and other columns contain the Mean_Abs_Err results of the other existing popular CF approaches. Having considered the values of Mean_Abs_Err from Table 5, the proposed USCS has low Mean_Abs_Err values than other existing CF approaches. It elucidates that the proposed USCS is comparatively more proficient, and optimal than other CF approaches. However, the proposed USCS for CF using WSD has some limitations. The USCS is partially searched based and rating independent. The integrated STC in the proposed USCS has ignored ratings of the old times though sometimes previously rated items should require recommending or suggesting to users because of enhancing the performance of predictions along with business demands.

5 Conclusion

Collaborative filtering can be considered as a strategically efficient recommendation system method that deals with appropriate manipulation of recommendations to users by predicting and analyzing their preferences. Determination of the user similarities has been enhanced the appropriateness and performance accuracy of the collaborative filtering approach. A comparatively proficient user similarity computation strategy for collaborative filtering has been represented that has overcome all the considerable challenges and become effective for predictions. The ADBM, WSD, STC, and SCM strategies have been integrated in the proposed USCS for CF. Integration of some other modules into it can also enhance the level of appropriateness for assumption.

Our future work is to deal with ratings of users provided to the old-rated items. Some old-rated items can be highly considerable for assuming the preferences, demands, requirements, and interests of users. Moreover, those old-rated items can be beneficial to recommend to the users commercially. In future, a desire is to predict a priority on those old-rated items that should be recommended to users. The method of prioritization will be free from starvation. Such predictions and prioritization will enhance the performance of recommendation system approach and fulfill the business demands completely.

References

1. Roy, F., Sheikh, M.-S., Mahamudul, H.: User similarity computation for collaborative filtering using dynamic implicit trust. In: International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham (2015)
2. Manolopoulus, Y., Nanopoulus, A., Papadopoulus, A.-N., Symeonidis, P.: Collaborative recommender systems: combining effectiveness and efficiency. Exp. Syst. Appl. **34**(4), 2995–3013 (2008)
3. Wang, H., Wang, N., Yeung, D.-Y.: Collaborative deep learning for recommender systems. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1235–1244 (2015)

4. Cacheda, F., Carneiro, V., Fern'andez, D., Formoso, V.: Comparison of collaborativefiltering algorithms: limitations of current techniques and proposals for scalable,high-performance recommender systems. In: ACM Transactions on the Web (TWEB), pp. 1–33 (2011)
5. Li, S., Kawale, J., Fu, Y.: Deep collaborative filtering via marginalized denoising auto-encoder. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management—CIKM'15, pp. 811–820 (2015)
6. Jin, Y., Hu, M., Singh, H., Rule, D., Berlyant, M., Xie, Z.: MySpace video recommendation with map-reduce on Qizmt. IEEE Internet Comput. INTERNET 126–133 (2010)
7. Zhao, Z.-D., Shang, M.-S.: User-based collaborative-filtering recommendation algorithms on Hadoop. In: Third International Conference on IEEE. Knowledge Discovery and Data Mining (2010)
8. Chen, Y.-Y., Cheng, A.-J., Winston, H.-H.: Travel recommendation by mining people attributes and travel group types from community-contributed photos. IEEE Trans. Multimedia **15**(6), 1283–1295 (2013)
9. Deng, S., Huang, L., Xu, G., Wu, X., Wu, Z.: On deep learning for trust-aware recommendations in social networks. IEEE Trans. Neural Networks Learn. Syst. **28**, 1164–1177 (2017)
10. Avesani, P., Massa, P., Tiella, R.: Moleskiing. it: a trust-aware recommender system for ski mountaineering. Int. J. Infonomics **20**, 1–10 (2005)
11. Homan, J.-V., Paul, J.-K.: A comparison of the relational database model and the associative database model. Issues Inf. Syst. **10**(1), 208–213 (2009)
12. McCarthy, D., John, C.: Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. Computat. Linguist. **29**(4), 639–654 (2003)
13. Lathia, N., Hailes, S., Capra, L.: Trust-based collaborative filtering. In: Karabulut, Y., Mitchell, J., Herrmann, P., Jensen, C.D. (eds.) Trust Management II, vol. 263, pp. 119–134. Springer, US (2008)
14. O'Donovan, J., Smyth, B.: Trust in recommender systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, pp. 167–174. ACM (2005)
15. Shambour, Q., Lu, J.: A trust-semantic fusion-based recommendation approach for e-business applications. Decis. Support Syst. **54**(1), 768–780 (2012)
16. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186. ACM (1994)
17. Bobadilla, J., Serradilla, F., Bernal, J.: A new collaborative filtering metric that improves the behavior of recommender systems. Knowl. Based Syst. **23**(6), 520–528 (2010)

Fast Predictive Artificial Neural Network Model Based on Multi-fidelity Sampling of Computational Fluid Dynamics Simulation

Monoranjan Debnath Rony, Mahmudul Islam, Md. Aminul Islam, and Mohammad Nasim Hasan

Abstract Computational cost associated with Computational Fluid Dynamics (CFD) simulation is a major bottleneck even with advanced computational facilities. This study introduces a method to overcome this major challenge by using multi-fidelity physics informed neural network (MPINN). Mesh domain elements in a CFD model determines the fidelity of our provided multi-fidelity analysis. CFD simulations with a lower number of mesh domain elements are computationally inexpensive but produce results with lower accuracy levels. In contrast, a large number of domain elements will result in higher accuracy with high computational cost. Using MPINN, we can accurately forecast the relationship between low and high-fidelity findings, which allows us to extract high accuracy CFD results using computationally cheap low-fidelity simulations. We conducted a simple stationary benchmark analysis of laminar mixed convective heat transfer in a square-shaped lid-driven cavity to find the optimal amount of high-fidelity training data needed for accurate outcomes while reducing computational costs. The findings reveal that the heat transfer performance parameter may be estimated with high precision while saving up to 51% in computing expenditures using MPINN with only 20% high-fidelity training data. This study is the first implementation of MPINN merging with CFD simulations to predict the performance of thermo-fluid systems. This method will allow us to solve complex industrial and engineering problems with higher accuracy while being computationally efficient. Nonetheless, this approach eventually led to faster system/product design, evaluation, and quality enhancement with resource optimization through the synergistic integration of multidisciplinary key components of the 4th industrial revolution.

M. D. Rony · M. Islam · Md. A. Islam · M. N. Hasan (✉)
Department of Mechanical Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka 1000, Bangladesh
e-mail: nasim@me.buet.ac.bd

M. Islam
e-mail: ridul@mit.edu

M. Islam
Department of Materials Science and Engineering, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_8

Keywords Computational Fluid Dynamics · Deep Neural Network · Multi-fidelity models · Resource optimization

1 Introduction

Computational Fluid Dynamics (CFD) simulation is a method that solves Navier–Stokes equations, as well as continuity and energy equations simultaneously, to simulate heat transfer and fluid flow properties in a system. The transport process involved in mixed convection heat transfer inside a vented cavity is a subject of great interest among the branches of CFD due to its vast use across numerous engineering disciplines and industrial applications. The forced flow created by the shear force by the sliding lid action is paired with the complex natural buoyancy-driven flow caused by the temperature gradient present in the differentially heated side walls. Digital electronic devices cooling [1], lubricating technology and chemical processing plants and equipment [2], food processing technologies [3], solar distiller [4], and other industrial and technical applications are all impacted heavily on heat transfer studies. Despite being such an extraordinary approach for predicting outcomes, CFD has some design constraints. Any CFD simulation method's one design restriction is that beyond the two end interval values of a certain parameter the simulation's accuracy and stability deteriorate. The size of the simulation system (number of mesh domain elements), timestep, and total simulation period are all the limitations in a CFD simulations. To acquire correct outcomes in an acceptable amount of time, we must set these parameters based on the designed problem statement and the available computational resources. The number of mesh domain elements should be kept to a minimum to ensure that the simulation is legitimate and conclusive. By increasing the solving point in a CFD problem, we can achieve more accurate solutions, but it comes with a penalty of substantial computational resources and time.

Several innovative methods have been studied to solve these design constraints of complex CFD problems. For enhancing the effectiveness and precision of CFD simulations, the Euler–Lagrange methodology paired with a discrete element method [5] has been investigated. The application of a higher stiffness coefficient matrix improves accuracy, and larger timesteps decrease computation time. As a result, within a certain number of domain decompositions, parallelization speeds up CFD simulations. A reduced-order analysis was proposed by Karcher and Wallraff [6], to predict unknown parameter combinations using Proper Orthogonal Decomposition with an interpolation method. For CFD solutions, they used Discontinuous Galerkin finite element method in combination with a nonlinear multigrid scheme. The purpose of our present study is to employ a multi-fidelity sampling-based artificial neural network (ANN) to reliably and efficiently forecast CFD simulation results.

In this modern era, astonishing advancement is being witnessed in the field of machine learning (ML) and neural network (NN) [7, 8]. With a sequence of algorithms that can reliably anticipate the link between particular kinds of data, a neural network can replicate the operation of brain function. These datasets are considered training data for the neural network. The accuracy of the training data determines the neural network's prediction performance. This is why incorporation between simulation-based computational physics and NN has been getting popular recently. The recent studies [9] reflect the combination of different fidelity simulation results, which saves computational time for generating NN training data. The studies [10, 11] show how these concepts were used to obtain improved accuracy and computing efficiency by pulling training data for NN from numerous models of various fidelity levels. Lu et al. [12] incorporated experimental and simulation data of nanoindentation of industrial alloys for training datasets and used a multi-fidelity neural network to learn and reduce systematic errors. Islam et al. [13] extracted material nanoscale properties accurately from molecular dynamics simulation using significantly less high-fidelity data saving a ton of computation expense. Meng and Karniadakis [14] constructed a multi-fidelity physics informed neural network (MPINN) that was carried out in this research. To considerably improve the prediction accuracy of NN, a huge number of computationally cheap low-grade simulation data is combined with a relatively small portion of precise and computationally demanding high-grade data.

Throughout this article, we used the MPINN to benchmark a popular CFD problem of mixed convection heat transfer in a square-shaped lid-driven cavity, and we were able to accurately forecast the result with considerably reduced computational expense.

2 Multi-fidelity Physics Informed Neural Network

The foundation of fidelity must be defined before any multi-fidelity investigation can begin. In CFD simulations, equations are solved numerically using various numerical methods and appropriate boundary conditions. For all numerical methods, at first, the entire domain must be divided into smaller elements that can be used to discretize the domain known as meshing. When the system is employed with less numbered mesh elements, the numerical simulation may not converge to give a solution. Even though the numerical simulation converges, the results are erroneous when compared to the simulations with a high numbered mesh elements system. But simulation using small number of mesh elements is computationally cheap, whereas high number of mesh elements give a highly accurate result, but they are computationally costly. We classified the findings from a small number of mesh element simulations as low-fidelity results and large number of mesh elements simulation resulted as high-fidelity data in our current multi-fidelity study.

The link between these low-grade and high-grade data is critical for any multi-fidelity modeling. This link was expressed as follows in a prior study [15]:

$$\mathrm{Nu_H} = \rho(x)\mathrm{Nu_L} + \delta(x) \tag{1}$$

where $\mathrm{Nu_H}$ and $\mathrm{Nu_L}$ represent low and high accuracy data, respectively, $\rho(x)$ represents the multiplicative correlation factor, and $\delta(x)$ represents the additive correlation factor. One of this relationship's flaws is that it can only manage linear correlation between two fidelity datasets, but in some other cases [16, 17] the relation between the low and high-grade data is nonlinear. For those cases, the relation between high- and low-fidelity data is expressed as

$$\mathrm{Nu_H} = F(x, \mathrm{Nu_L}) \tag{2}$$

$F(.)$ is an unknown (nonlinear/linear) function that correlates low- to high-fidelity data, which was investigated by Meng and Karniadakis [14]. They divided $F(.)$ into nonlinear and linear components, which are written as follows:

$$F = F_\mathrm{l} + F_\mathrm{nl} \tag{3}$$

where the linear and nonlinear factors in F are F_l and F_nl, respectively. As a result, the relationship between low-fidelity and high-fidelity data is as follows:

$$\mathrm{Nu_H} = \alpha F_\mathrm{l}(x, \mathrm{Nu_L}) + (1-\alpha) F_\mathrm{nl}(x, \mathrm{Nu_L}), a \in [0, 1] \tag{4}$$

where α is a hyper-parameter that specifies the degree of nonlinearity between low-fidelity and high-fidelity data, and the MPINN must be trained to determine the linear–nonlinear relationships, as well as the hyper-parameter.

Meng and Karniadakis [14] designed and validated the MPINN architecture used for this investigation. Three fully connected neural networks form the MPINN, as shown in Fig. 1. The first one ($\mathrm{NN_L}$) approximates low-fidelity data while the second ($\mathrm{NN_{H1}}$) and third ($\mathrm{NN_{H2}}$), together denoted as $\mathrm{NN_H}$, estimate the linear (F_l) and nonlinear (F_nl) correlations between the low- and high-grade data, respectively. The predictive accuracy of MPINN is strongly influenced by its size (depth and width). As there is so much low-fidelity data, finding the right size for $\mathrm{NN_L}$ to resemble a low-fidelity function is straightforward. However, few numbers of high-fidelity data are available due to its computational expense. So, importance should be given to the size of $\mathrm{NN_H}$ while designing MPINN. Meng and Karniadakis [14] proposed the best range for $\mathrm{NN_{H2}}$ with depth (l) and width (w) as $l \in [1,2]$ and $w \in [4,20]$. In our investigation, $\mathrm{NN_{H2}}$ of our MPINN is comprised of two hidden layers ($l = 2$), each with 20 neurons ($w = 20$). As only the linear link between low and high-fidelity data is predicted by $\mathrm{NN_{H1}}$. As a result, only one hidden layer with one neuron is recommended. The neural network's specifications are as follows (Table 1).

The loss function that has been optimized in the present study is

$$\mathrm{MSE} = \frac{1}{N_{\mathrm{Nu_L}}} \sum_{i=1}^{N_{\mathrm{Nu,L}}} (\mathrm{Nu_L^*} - \mathrm{Nu_L})^2 + \frac{1}{N_{\mathrm{Nu_H}}} \sum_{i=1}^{N_{\mathrm{Nu,H}}} (\mathrm{Nu_H^*} - \mathrm{Nu_H})^2 \tag{5}$$

where $\mathrm{Nu_L^*}$ and $\mathrm{Nu_H^*}$ denote the output of the $\mathrm{NN_L}$ and $\mathrm{NN_H}$, respectively. $\mathrm{NN_{H2}}$ may yield overfitting as a small number of high-fidelity data is available, which is why to avoid overfitting, we have used $\mathrm{L_2}$ regularization only in $\mathrm{NN_{H2}}$ neural network.

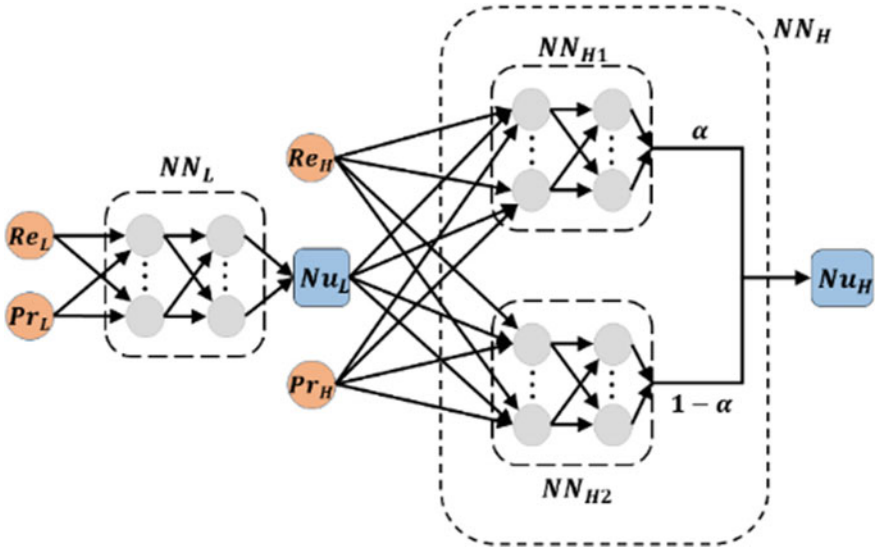


Fig. 1 Multi-fidelity physics informed neural network (MPINN) implemented in this study

Table 1 Detailed information of the MPINN's neural networks

Neural network	Number of hidden layers	Neurons per hidden layer	Activation function	L_2 regularization parameter
NN_L	4	20	Input and hidden layers: tangent sigmoid function Output layers: linear function	0
NN_{H1}	1	1	Linear function	0
NN_{H2}	2	20	Tangent sigmoid function	0.001

The regularization parameter in this neural network is set to 0.001, which has been determined to prevent overfitting and assure no substantial underfitting across a diverse variety of datasets. The loss function is optimized in this work utilizing the L-BFGS approach [18] combined along with Xavier's initialization method [19]. Reynolds number (Re) and Prandtl number (Pr) are the design input parameters (x) of the MPINN in this work, and Nusselt number (Nu) is the target output that the MPINN is taught to predict accurately. In this study, the MPINN is used in MATLAB [20].

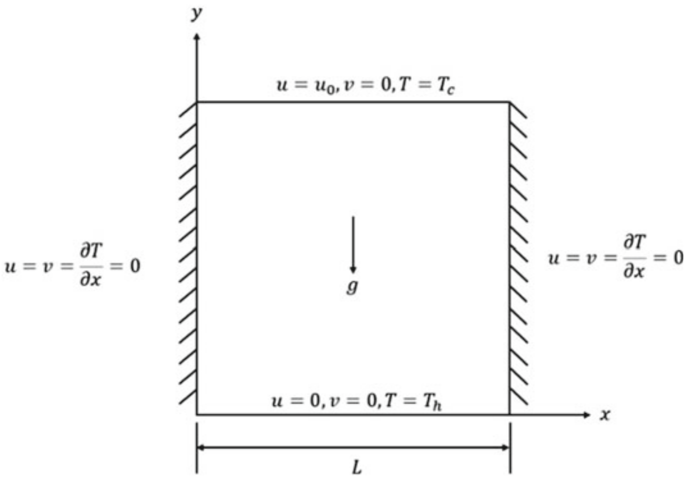


Fig. 2 Schematic diagram of the physical domain

3 Benchmark Study

3.1 Physical Model of Benchmark CFD Simulation

The benchmark study of CFD is a mixed convection heat transfer problem. Figure 2 depicts the study's physical domain, which is a square-shaped vented cavity of length *L*. The upper horizontal lid moves at a constant velocity rate of u_0, while the lower lid remains motionless. The enclosure's top and bottom walls are considered to be at a constant lower (T_c) and higher (T_h) temperature, respectively, and the vertical walls remains adiabatic.

3.2 Mathematical Model of Benchmark CFD Simulation

Like most CFD problems, this simulation is performed considering several assumptions. Laminar, incompressible, Newtonian, steady-state, and two-dimensional fluid motion is assumed inside the chamber. The Boussinesq approximation, which connects flow and thermal fields, is used to handle changes in fluid density with temperature. The rest of the fluid's thermophysical properties are temperature independent. All fluid interfaces that come into contact with the solid have no-slip velocity boundary conditions. The effects of thermal radiation, Joule heating, and viscous dissipation are not included in the energy calculation because they have very less effect on the output.

The Navier–Stokes equations, as well as the continuity and energy equations utilized in this study, were normalized and non-dimensioned to account for these considerations and presented as follows:

$$\frac{\partial U}{\partial X}+\frac{\partial V}{\partial Y}=0 \tag{6}$$

$$\frac{\partial U}{\partial \tau}+U\frac{\partial U}{\partial X}+V\frac{\partial U}{\partial Y}=-\frac{\partial P}{\partial X}+\frac{1}{\mathrm{Re}}\left(\frac{\partial^2 U}{\partial X^2}+\frac{\partial^2 U}{\partial Y^2}\right) \tag{7}$$

$$\frac{\partial V}{\partial \tau}+U\frac{\partial V}{\partial X}+V\frac{\partial V}{\partial Y}=-\frac{\partial P}{\partial Y}+\frac{1}{\mathrm{Re}}\left(\frac{\partial^2 V}{\partial X^2}+\frac{\partial^2 V}{\partial Y^2}\right)+\mathrm{Ri}\theta \tag{8}$$

$$\frac{\partial T}{\partial \tau}+U\frac{\partial T}{\partial X}+V\frac{\partial T}{\partial Y}=\frac{1}{\mathrm{RePr}}\left(\frac{\partial^2 \theta}{\partial X^2}+\frac{\partial^2 \theta}{\partial Y^2}\right) \tag{9}$$

where U and V are the X- and Y-direction velocity components, respectively, θ is the temperature, and P is the pressure.

The Reynolds number, Grashof number, Prandtl number, and Richardson number are the non-dimensional numbers Re, Gr, Pr, and Ri, which are defined as

$$\mathrm{Re}=\frac{\rho v_{\max}H}{\mu},\ \mathrm{Gr}=\frac{g\beta(T-T_{\mathrm{c}})H^3\rho^2}{\mu^2},\ \mathrm{Ri}=\frac{\mathrm{Gr}}{\mathrm{Re}^2},\ \mathrm{Pr}=\frac{\mu}{\rho\alpha}$$

For getting the normalized form of the aforementioned governing Eqs. (6)–(9), the non-dimensional procedure is performed using the dimensionless parameters listed below.

$$X=\frac{x}{H},\ Y=\frac{y}{H},\ U=\frac{u}{v_{\max}},\ V=\frac{v}{v_{\max}},$$
$$P=\frac{p-p_{\mathrm{o}}}{\rho v_{\max}^2},\ \theta=\frac{T-T_{\mathrm{c}}}{T_{\mathrm{h}}-T_{\mathrm{c}}},\ \tau=\frac{t v_{\max}}{H}$$

where the fluid density, coefficient of volumetric expansion, kinematic viscosity, thermal diffusivity, and gravitational acceleration are expressed by ρ, β, υ, α, and g, respectively.

The thermal and velocity boundary conditions in the dimensionless form corresponding to the current numerical model are described in Table 2.

Table 2 Boundary conditions in non-dimensional form

Boundary	Thermal field	Velocity field
Top wall	$\theta=0$	$U=1,\ V=0$
Bottom wall	$\theta=1$	$U=V=0$
Vertical walls	$\partial\theta/\partial X=0$	$U=V=0$

The system performance representing the steady heat transfer rate from the hot bottom wall is evaluated as Average Nusselt Number (Nu) as defined below:

$$\mathrm{Nu}_{\mathrm{avg}} = -\int_0^1 \left(\frac{\partial \theta}{\partial Y}\right)_{Y=0} \mathrm{d}X \tag{10}$$

3.3 Simulation Procedure of Benchmark Study

A finite element formulation based on the Galerkin weighted residual technique is used to model and solve the governing equations from (6) to (9) and the boundary conditions from Table 2. A nonlinear parametrial solver is used to resolve the governing equations, ensuring faster convergence and dependability. Non-uniform quadrilateral mesh elements are used in the current experiment.

We used a homogeneous grid of Re and Pr to generate performance attributes under various system circumstances. The parameters of this MPINN analysis for the CFD model are Re, which ranges from 100 to 1000 at 100 intervals, and Pr, which ranges from 0.1 to 7.1 at 1 interval. The entire uniform input parameter sample space for the CFD system consists of 80 sample points. The system with a mesh consisting of 350 domain elements and 48 boundary elements (Fig. 3) generates low-fidelity training data for MPINN, whereas the mesh, which consists of 26,310 domain elements and 600 boundary elements (Fig. 4) generates high-fidelity data.

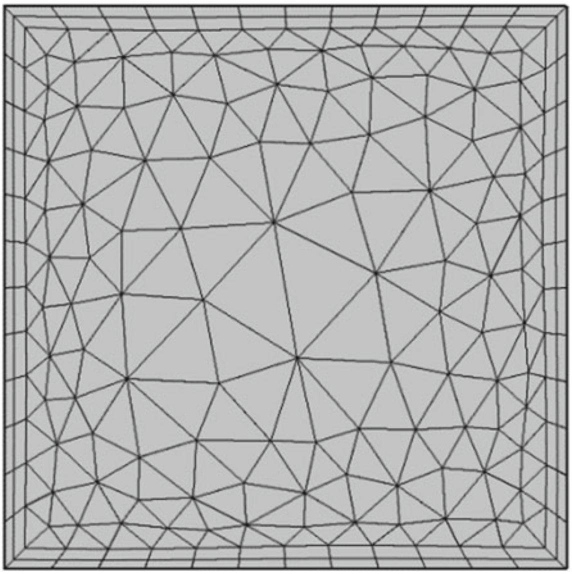


Fig. 3 Mesh distribution for low-fidelity system

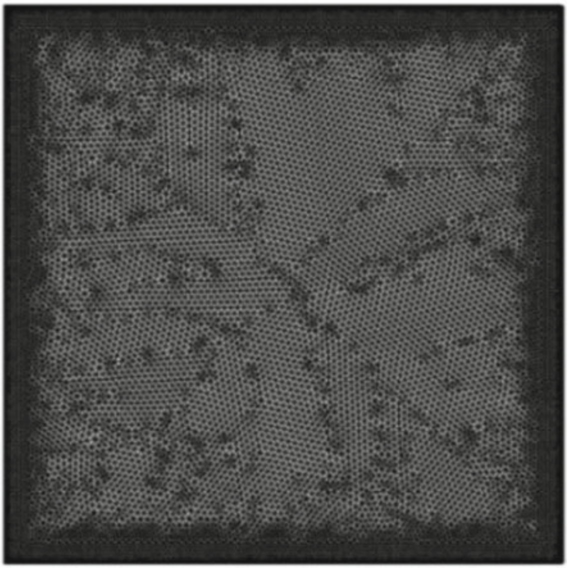


Fig. 4 Mesh distribution for high-fidelity system

Although the proposed MPINN was meant to accurately forecast a wide sample space with a small number of high-fidelity samples, an equal number of high-fidelity and low-fidelity simulations were run for our present benchmark research. The following equation of mean absolute percentage error (MAPE) is used to calculate all of the errors in the current study.

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{A_i - F_i}{A_i}\right| \tag{11}$$

Here, A_i is the high-fidelity CFD simulation's system property value, and F_i is the MPINN or low-fidelity CFD simulation's system property value. As this is a benchmark study, 10 different initialization parameters were used to train the MPINN for each of the examples examined. The MAPE values used in this study's plots are the median values of MAPE among these 10 MPINN prediction errors.

4 Results and Discussion

The benefit of using MPINN method is that we do not need the same number of high- and low-fidelity training data. Very few numbers of more accurate and reliable, computational costly high-fidelity data along with many low-fidelity less accurate and easily attainable training data combinedly predicts the output accurately with huge computational savings. There is total 80 low-fidelity data for the study, the same as the number of sample points. In every scenario, MPINN has been trained using these 80 low-fidelity samples. If we employ 8 high-fidelity data points from the uniform sample field in addition to the low-fidelity input, the method is presented as a trained MPINN with 10% high-fidelity data. We used a wide variety of high-fidelity sampling (10–90%) to train our MPINN to evaluate the forecasting performance of any CFD simulations in this benchmark study. We have evaluated and predicted Nusselt Number (Nu). Figure 5 shows the mean absolute percentage errors (MAPE) of the MPINN in predicting the system's performance parameter (Nu) for various percentages of high-fidelity training samples. It can be seen that the MAPE reduces as the MPINN is trained with higher % of high-fidelity sampling. With improved quality and higher number of training data, our MPINN, like most deep neural networks, can predict more reliably. The MAPE for Nu is 1.86% when 10% high-fidelity training set is present. For 20% high-fidelity data, it drops to 0.55%, and for 30% high-fidelity data, it drops to 0.44%. After that, decrement is very small, nearly around 0.2%; meanwhile, they are becoming computationally heavy. If MAPE is smaller than 1–3%, then we are convinced that the prediction is very good.

As a result, we may deduce that only 20% high-fidelity sampling data is sufficient to lower the MAPE to 0.55%, resulting in a computational savings of more than 50%.

Therefore, with only 20% of high-fidelity training data, we can achieve very accurate predictions and reduce the prediction error significantly. The contours in Fig. 6 illustrate a valid comparison between the high-fidelity modeling, low-fidelity modeling, and the output by the trained MPINN using 20% high-fidelity data. Figure 6a, b shows that the contours of high and poor fidelity are considerably distinct, particularly in the region where Pr is low and Re is high. The MAPE for the low-fidelity simulation is 5.88%, and the addition of only 20% high-fidelity data

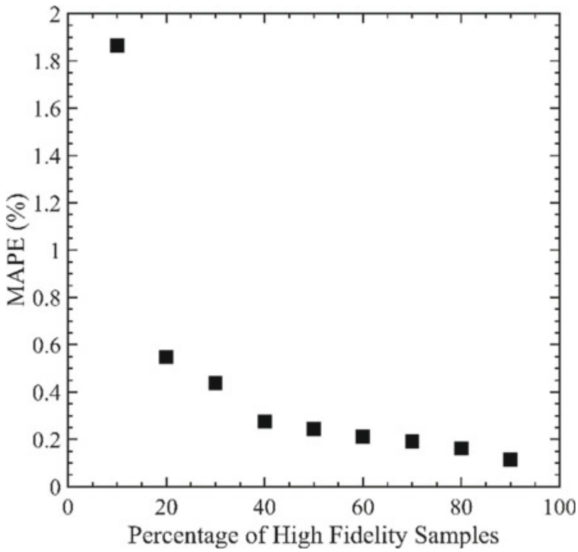


Fig. 5 MAPE of MPINN predicting Nu for different high-fidelity sample percentages

results in 10 times reduction of the MAPE, to a value of only 0.55%, Fig. 6c contour confirms that. The contour of high-fidelity simulation and 20% high-fidelity data trained MPINN system is practically identical.

When compared to traditional CFD simulations, MPINN was capable of achieving high precision CFD simulation which results with significantly lower computing cost in our current investigation. The computational savings (CS) have been calculated using the following Eq.:

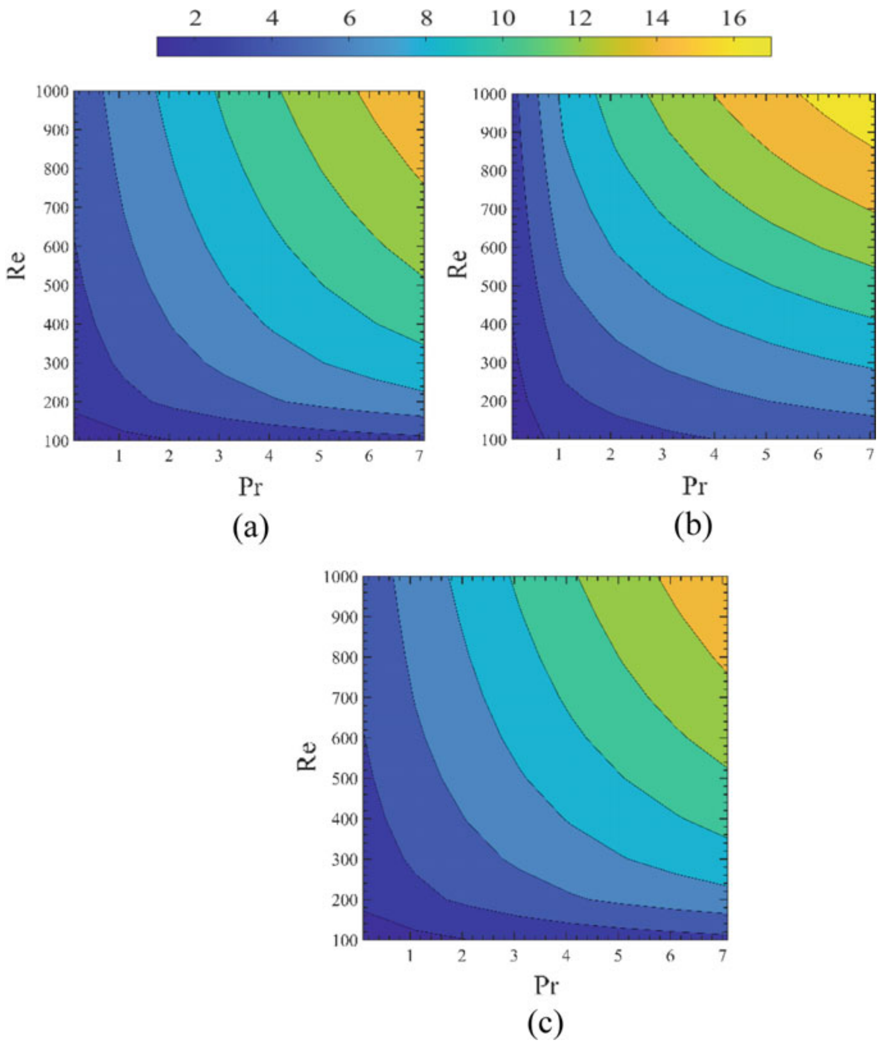


Fig. 6 Contours of Nusselt Number of the benchmark study acquired from (**a**) high-fidelity modeling, (**b**) low-fidelity modeling (MAPE = 5.88%), and (**c**) 20% high-fidelity modeled MPINN (MAPE = 0.55%)

$$\text{CS} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\text{CC}_{\text{HF}} - \text{CC}_{\text{LF}}}{\text{CC}_{\text{HF}}}\right| \tag{12}$$

where CC_{HF} and CC_{LF} denote the computing expense of high and low-fidelity CFD simulations, respectively. Both low- and high-fidelity CFD simulations were performed using a workstation consisting of a processor of 2.5 GHz clock speed, 8 GB RAM, and 64-bit operating system.

From Fig. 7, it can be clearly seen when CS is maximum, MAPE is quite high. But when CS is reduced from 61% to 51%, MAPE also drops from 1.86% to 0.55%, which is a 70% reduction in MAPE. Further, lowering CS from 51% down to 41% reflected with a 20% reduction in MAPE. Computational savings is inversely proportional to the percentage of high-fidelity data. Computational savings drop only 10% from 61% to 51% when the percentage of high-fidelity sample is increased from 10% to 20%. As previously stated, using only 10% high-fidelity data reduces MAPE from 5.86 to 1.86%, which is a 68% reduction in MAPE. Further, increase in percentage high-fidelity data from 10% to 20% decreases MAPE by 70%. At the same time, computational savings increase only 10%. Further, increase in percentage high-fidelity data from 20% to 30% has less effect in decreasing MAPE while the reduction in CS remains as before.

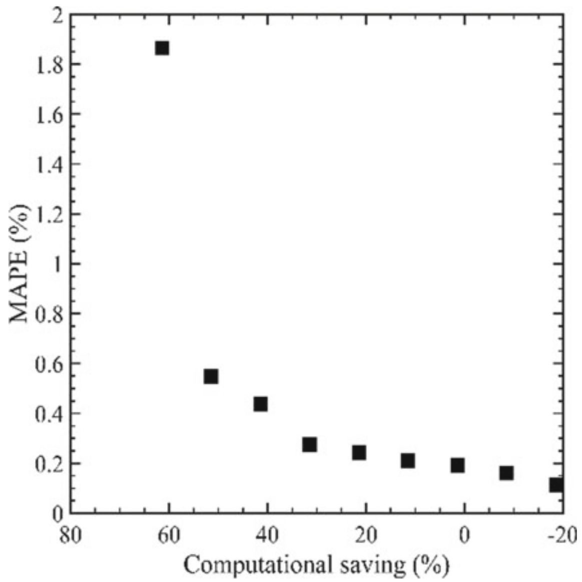


Fig. 7 MAPE(%) versus computational savings (%) in predicting Nu using MPINN

5 Conclusion

In this paper, we explained how to use a multi-fidelity physics informed neural network (MPINN) to lower the computing expense of a computational fluid dynamics (CFD) simulation. The amount of mesh elements determines the simulation fidelity. The trends of representative system performance in terms of Average Nusselt Number using only 20% high-fidelity training data can be accurately forecast using the developed MPINN. It's also seen that, the trends of our two different fidelity data could be quite different, yet the precision of MPINN can be quite of high accuracy. MPINN's high accuracy predictions are accompanied by a processing time reduction of 51%. In addition, regardless of the complexity of high-fidelity simulations, the computational expense of training MPINN is quite low. The current approach essentially leads to faster system/product design, assessment, and subsequent quality improvement with optimum resource utilization through the synergistic integration of multidisciplinary key components of the 4th industrial revolution.

Acknowledgements This research was carried out at the Mechanical Engineering Department, Bangladesh, University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh.

References

1. Nimmagadda, R., Reuven, R., Asirvatham, L.G., Wongwises, S.: Thermal management of electronic devices using gold and carbon nanofluids in a lid-driven square cavity under the effect of variety of magnetic fields. IEEE Trans. Compon. Packag. Manuf. Technol. **10**(11), 1868–1878 (2020). https://doi.org/10.1109/tcpmt.2020.3008786
2. Billah, M., Rahman, M., Sharif, U.M., Rahim, N., Saidur, R., Hasanuzzaman, M.: Numerical analysis of fluid flow due to mixed convection in a lid-driven cavity having a heated circular hollow cylinder. Int. Commun. Heat Mass Transfer **38**, 1093–1103 (2011)
3. Sun, D.-W.: Thermal Food Processing: New Technologies and Quality Issues, 2nd edn. CRC Press (2012)
4. Kolsi, L.: Heat and mass transfer in 3D inclined lid-driven solar distiller. Int. J. Fluid Mech. Therm. Sci. **1**, 72 (2016)
5. Alobaid, F., Baraki, N., Epple, B.: Investigation into improving the efficiency and accuracy of CFD/DEM simulations. Particuology **16**, 41–53 (2014)
6. Karcher, N., Wallraff, M.: Accelerating CFD solver computation time with reduced-order modeling in a multigrid environment. Int. J. Numer. Meth. Fluids **93**(2), 462–480 (2020). https://doi.org/10.1002/fld.4892
7. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016). https://doi.org/10.1038/nature16961
8. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D.: Mastering the game of go without human knowledge. Nature **550**(7676), 354–359 (2017). https://doi.org/10.1038/nature24270

9. Pilania, G., Gubernatis, J., Lookman, T.: Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput. Mater. Sci. **129**, 156–163 (2017). https://doi.org/10.1016/j.commatsci.2016.12.004
10. Kennedy, M.C., O'Hagan, A.: Predicting the output from a complex computer code when fast approximations are available. Biometrika **87**, 1–13 (2000)
11. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Inferring solutions of differential equations using noisy multi-fidelity data. J. Comput. Phys. **335**, 736–746 (2017). https://doi.org/10.1016/j.jcp.2017.01.060
12. Lu, L., Dao, M., Kumar, P., Ramamurty, U., Karniadakis, G.E., Suresh, S.: Extraction of mechanical properties of materials through deep learning from instrumented indentation. Proc. Natl. Acad. Sci. **117**(13), 7052–7062 (2020). https://doi.org/10.1073/pnas.1922210117
13. Islam, M., Thakur, M.S., Mojumder, S., Hasan, M.N.: Extraction of material properties through multi-fidelity deep learning from molecular dynamics simulation. Comput. Mater. Sci. **188**, 110187 (2021). https://doi.org/10.1016/j.commatsci.2020.110187
14. Meng, X., Karniadakis, G.E.: A composite neural network that learns from multi-fidelity data: application to function approximation and inverse PDE problems. J. Comput. Phys. **401**, 109020 (2020). https://doi.org/10.1016/j.jcp.2019.109020
15. Fernández-Godino, M.G., Park, C., Kim, N.-H., Haftka, R.T.: Review of multi-fidelity models. (2016)
16. Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N.D., Karniadakis, G.E.: Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. Proc. R. Soc. A: Math. Phys. Eng. Sci. **473**, 20160751 (2017). https://doi.org/10.1098/rspa.2016.0751
17. Babaee, H., Perdikaris, P., Chryssostomidis, C., Karniadakis, G.E.: Multi-fidelity modelling of mixed convection based on experimental correlations and numerical simulations. J. Fluid Mech. **809**, 895–917 (2016). https://doi.org/10.1017/jfm.2016.718
18. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**, 503–528 (1989). https://doi.org/10.1007/BF01589116
19. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) JMLR Workshop and Conference Proceedings, Chia Laguna Resort, Sardinia, Italy, pp. 249–256 (2010). http://proceedings.mlr.press/v9/glorot10a.html
20. MATLAB, 9.7.0.1190202 (R2019b), The MathWorks Inc., Natick, Massachusetts (2018)

Data and Network Security

Reducing Product Counterfeiting Using Blockchain Technology in E-Commerce Business

Md. Rashed Bhuiyan, Mohammod Abul Kashem, Fahmida Akter, and Salma Parvin

Abstract Due to the lack of traceability and proper authentication process, product counterfeiting is one of the main challenges in supply-chain management. In e-commerce business, merchants act as the central and only single controlling authority which leads to less trust and transparency. This creates a chance of product counterfeiting which can be reduced using blockchain with other supporting technologies. As products are delivered in multiple channels through many hands, there needs to be a way to identify original products and a proper product handover process. Many technologies are introduced to combat product counterfeiting such as barcode, QR code, NFC tag, hologram, etc. But these technologies are either expensive or lack full proof of product counterfeiting. Our research shows that a combination of destructible dynamic QR code, product location tracking along with blockchain technology can greatly reduce product counterfeiting in the e-commerce business. This paper proposes a novel approach of a product authentication system where a blockchain-based web application and a mobile application are introduced instead of a traditional product authentication system. Our evaluation shows that using this system, a consumer can easily identify whether a product is counterfeit or not.

Keywords Counterfeiting · Blockchain · E-commerce · Traceability · QR code

Md. R. Bhuiyan (✉) · M. A. Kashem
Dhaka University of Engineering and Technology, Gazipur, Dhaka, Bangladesh
e-mail: emailofrashed@gmail.com

M. A. Kashem
e-mail: drkasheml1@duet.ac.bd

F. Akter
Bangladesh University of Professionals, Dhaka, Bangladesh

S. Parvin
Jahangirnagar University, Savar, Dhaka, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_9

1 Introduction

The selling and purchasing of products and services through the Internet is known as e-commerce. With the growing usage of e-commerce platforms as a key engine of online sales and financial transactions, protecting consumers' rights and the brand value of reputable firms has become a big problem [1–4]. Counterfeiters and pirates create and distribute goods that are frequently inferior and, in some cases, hazardous, posing health and safety concerns ranging from minor to life-threatening [5, 6]. Product counterfeiting also causes environmental hazards, loss of sales revenue and profit for the business, warranty claims for fraudulent products, damage of brand and products reputation and loss of money as consumers. Almost any product can be counterfeited, and even experts might have difficulty distinguishing authentic from counterfeit products. Moreover, because of the current size and high growth rate of e-commerce business, getting consumers' trust is an issue nowadays. In 2016, the global trade in counterfeit and pirated goods reached 509 billion US dollar, accounting for 3.3% of global trade, and the number is increasing day by day [7]. Third-party logistics services are the most popular way of shipping counterfeit products, and many e-commerce businesses use that [8]. As a result, business organizations spend a significant amount of money and time protecting their brands. Despite all of the safeguards, the counterfeit business continues to grow at a rapid pace.

One of the main problems of e-commerce business is the lack of transparency [9]. Consumers can't easily identify whether the products they buy online through e-commerce websites are counterfeit or genuine. The fact that the product is being delivered to the consumer through a few channels can lead to errors in product delivery, mismatch of one product with another, or the possibility of counterfeiting. The consumers cannot genuinely trace every step of product handling from each delivery channel rather they solely depend on the merchants' notifications. As e-commerce operates through websites, the consumers have very low visibility of the products and it is easy to deceive the consumers because there is no way to check the authenticity of the product before delivery [10, 11]. Besides, the consumers and the sellers are not familiar with each other, and they reside in different locations. Recently, in Bangladesh some e-commerce websites such as Evaly, eorange, and Dhamaka have made news headlines for cheating with customers. Even customers' information were deleted from databases to cover up their misdeeds. As a result of this dishonest activities by a segment of e-commerce websites, a significant trust deficit has been formed between customers and e-commerce. The COVID-19 pandemic has also provided opportunities to the criminals to sell more counterfeit products online as consumers are staying home and purchasing through e-commerce websites more than before [12]. The centralized e-commerce database has the risk of system failure, and it is vulnerable to potential attacks on the system by malicious participants [13, 14]. It is also possible for merchants to perform modifications of product or consumer order records without leaving any trace for future investigation. To combat the growing concern, it is necessary now to ensure innovative, affordable, full proof, user-friendly, and easy-to-implement product anti-counterfeiting and traceability solutions in the e-commerce industry.

To add transparency, accountability, traceability, and trustworthiness in every level of product handling in the e-commerce business, we have developed a prototype web application and a mobile application based on destructible dynamic QR code, location tracking and blockchain technology that can be used for checking the authenticity of a product by scanning QR code using appropriate credentials. As the system will generate much data against a product, relevant stakeholders will be able to utilize these data for sales and marketing too.

1.1 Supply-Chain Management of an E-Commerce

Generally, an e-commerce company operates a website, and that website sells various products to consumers through multiple steps. We call him a merchant who owns an e-commerce website. When a product order is placed by a consumer, the warehouse receives the order through the merchant and the responsible person of the warehouse ensures the product is ready for delivery. The order is then delivered to the shipping provider from the warehouse. The shipping company may have its own delivery hubs in different locations. After a package is collected from the consumer's nearest hub by shipping, the package arrives at the consumer's door and the consumer receives it. Once the consumer receives the product, this completes the cycle of an order (Fig. 1).

Fig. 1 A typical supply chain for an e-commerce system

1.2 Destructible Dynamic QR Code

A destructible QR code is more secure than a normal paper-printed QR code. Metalized vinyl is used to print the destructible labels of the QR codes. They are impossible to remove in one piece and will disintegrate if someone tries. Thus, removing a QR code from one product and sticking it to the other product is not possible. This is a barrier to product counterfeiting. The QR code can be dynamic as well as destructible. A short unique URL only for the product is printed on the QR code. A dynamic QR code is similar to a static QR code but contains a short redirection URL. Static QR code is not editable, and using this QR code, information cannot be tracked [15]. Whereas the dynamic QR code does not store complete information, rather it stores a short URL that redirects the user to the target information located in the web server. As the QR code is dynamic, it can present new information from time to time managed from the web server. Besides, the dynamic QR code does not contain actual messages, rather only a short URL assigned to it. Due to this, the QR code is much simpler and smaller in size. This makes the advantage of quicker and easier scan because there are fewer characters in it.

1.3 Blockchain

Blockchain is a shared, distributed and tamper-proof ledger for recording transactions [16]. Blocks are used to store transactions in the blockchain. A cryptographic hash function is used to link each block in the chain to the previous block. Whenever a new transaction happens on the blockchain, a record of it is written to each participant's ledger. This implies that if a single block in a chain is altered, it will be immediately evident that the chain has been tampered with. Malicious participants or hackers would have to modify every block in the chain, across all distributed copies of the chain, if they intended to damage a blockchain system [17]. Initially, it was made for financial transactions, but it is now revolutionizing the business world especially e-commerce, and many applications have emerged from this technology [18, 19]. Any application using this technology as its base architecture ensures that the contents of its data are protected and transparency at all levels is maintained. There exist significant applications based on blockchain technology, and enormous applications are gradually being developed. There are mainly four categories of blockchain—public, private, consortium, and hybrid blockchain [20]. Both the public and private blockchain networks can be permissioned. This limits who is permitted to join in the network and what transactions they may do. To participate, participants must first receive an invitation or authorization.

2 Related Works

A good number of research papers have been published regarding the technologies and techniques of product counterfeiting, but most of them are either expensive, have security drawbacks, or do not solve problems according to every user's needs. Such as, Choi and Poon [21] proposed a track-and-trace anti-counterfeiting system using RFID which was targeted for high-end consumer products using a central database. Chen and Chen [22] proposed a QR code-based product authentication system using a digital certificate and digital signature which has multiple QR code printing issues and security vulnerabilities including QR code tampering. Khalil et al. [23] introduced an RFID-based scheme for anti-counterfeiting in large-scale retail environments but using a centralized server and no scope of counterfeit location tracking. Authors in [24] showed an IoT-based anti-counterfeiting system with the combination of RFID and QR code for prepackaged food items which was not suitable for all e-commerce products and the system lacks product location traceability. Authors in [25] explained how blockchain technology can be used in the pharmaceutical industry to prevent counterfeit drugs, but there was no mention of location and scanning data tracking. Authors in [26] proposed textile and clothing product authentication technology but with static QR codes which were integrated in the clothing item multiple times in multiple stages, unchangeable once printed and can be easily copied. Authors in [27] presented a NFC-based framework to detect consumer-level counterfeit but till now most of the consumers do not have NFC mobile phones.

Compared to the solutions of different researches, the system we have proposed here has some advantages. We have proposed a system where there is no expensive technology is introduced like the NFC or RFID. Also, we have used dynamic QR codes, whereas most of the solutions provide static QR codes. Nobody proposed a destructible feature of the QR code but we did. Tracing every step in the supply chain is a must for a full-proof system from the seller to the consumer, and we have proposed the method like this way, whereas most of the methods proposed partial tracing.

3 Methodology

Our proposed model consists of a web and a mobile application. The web application is used by the merchant, the warehouse, the delivery hub, and the shipping. The mobile application is used by the consumers. Each user needs to register first before using the system. A location access permission will be asked before using the application, and allowing it will give access to the system using a valid username and password. When a consumer orders for a product, a unique URL will be generated against that product, and it will be maintained in the whole journey of the product delivery. The unique product URL is generated randomly except the domain name. Only the content of the redirection page will be changed based on the user who is

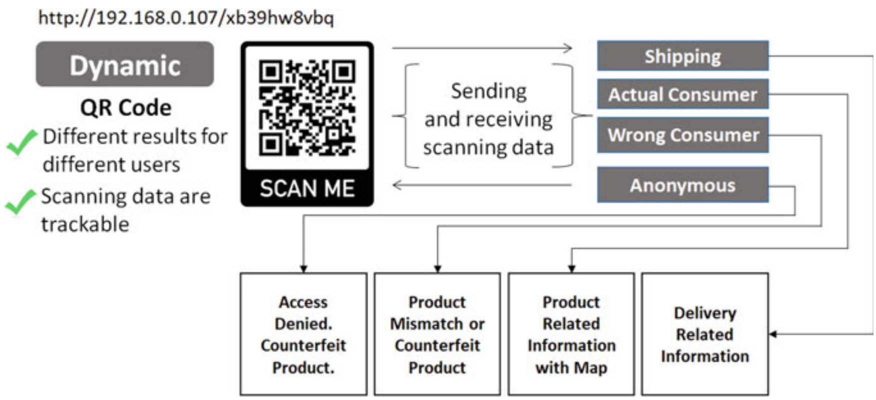


Fig. 2 Different scanning results for different users

scanning the QR code. The QR code will be printed from the warehouse with the generated unique URL. A destructible dynamic QR code is used so that no malicious participant can remove the QR code sticker from the product and data can be tracked, e.g., user information, GPS location, device information, date-time, number of scanning when a scan happens. The dynamic QR code is used as a unique identifier specific to a product. This QR code is stuck to the product if possible otherwise on the product package (Fig. 2).

During the transportation process of the product from warehouse to consumer, each involved user scans the QR code and the scanned data are added into the blockchain through the unique product number respectively. In this way, the next user can check whether or not the product has already passed through the valid product provider and the product-related information is valid or not. If there is any inconsistency found after scanning, the product may be considered counterfeit and should be returned to the current product provider.

The following steps are followed during product delivery from the warehouse to the consumer:

Step 1: The warehouse prints the QR code and sticks it to the product.
Step 2: Shipping receives the product, scans the QR code, and takes or returns the product.
Step 3: Shipping drops the product at the consumer's nearest delivery hub.
Step 4: The delivery hub receives the product, scans the QR code, and takes or returns the product.
Step 5: Shipping receives the product, scans the QR code, and takes or returns the product.
Step 6: Consumer receives the product, scans the QR code, and takes or returns the product.

User authentication is made before scanning the QR code. Thus any unregistered user beyond the system is not allowed to scan the QR code. If any unregistered user

Algorithm 1 Product Authentication by Scanning QR Code

Require: User's public key U_{pk}, Unique product number P_{num}
Ensure: Scan result S_r

function AUTHENTICATION (U_{pk}, P_{num})
  $U_t \leftarrow$ find user type (U_{pk}, P_{num})
  if U_t = actual user **then**
    $U_d \leftarrow$ Get unique product code with location and device information
    Add a block of this scanning transaction into the blockchain (U_d)
    $S_r \leftarrow$ Show product related information
  else if U_t = registered but wrong user then **then**
    $U_d \leftarrow$ Get unique product code with location and device information
    Add a block of this scanning transaction into the blockchain (U_d)
    $S_r \leftarrow$ Show product mismatch or counterfeit message
  else
    $U_d \leftarrow$ Get unique product code with location and device information
    Add a block of this scanning transaction into the blockchain (U_d)
    $S_r \leftarrow$ Show access denied and counterfeit message
  end if
end function

scans the QR code using a third-party app, i.e., a scan happens by an anonymous, he will not be able to see product-related info. Instead, a product counterfeit message will be shown, and his device details, location, and scanning information will be tracked to find counterfeit products. Though using the proposed product authentication system, product mismatch will not be possible, but it may happen due to the biological nature of humans. That's why if a registered user who didn't place the order receives a wrong product mistakenly during product handover and the user scans the product QR code, the scanning will show product mismatch or counterfeit product message. The consumer will scan the QR code of the ordered product through the provided android application. The scanning will show details of the product with related information such as where the product is scanned, how many times the product is scanned, location path of the product journey from warehouse to consumer, etc. When the consumer scans the QR code after delivery, if there is no anomaly the product can be recorded as sold in the blockchain. Google Map API [28] is used here to draw paths. When the responsible person of the warehouse enters data from his account, his GPS location is tracked to draw the starting point of the product journey path. In the same way, then the shipping's location and the nearest delivery hub's GPS location are tracked, and when the consumer receives the product, his GPS location is also tracked. This will create a product journey map, and it is important for the consumer to detect any unusual product handover (Fig. 3).

Private and permissioned blockchain is used in this system as this is involved with the supply-chain system, and the transaction details are only shared with the concerned parties [29, 30]. The private blockchain is lighter than the public blockchain and implementation cost is very low. As the blockchain data is smaller and there are fewer transactions than public blockchain, private blockchain can act faster. In our private blockchain, it is easier to identify users since each user is provided with

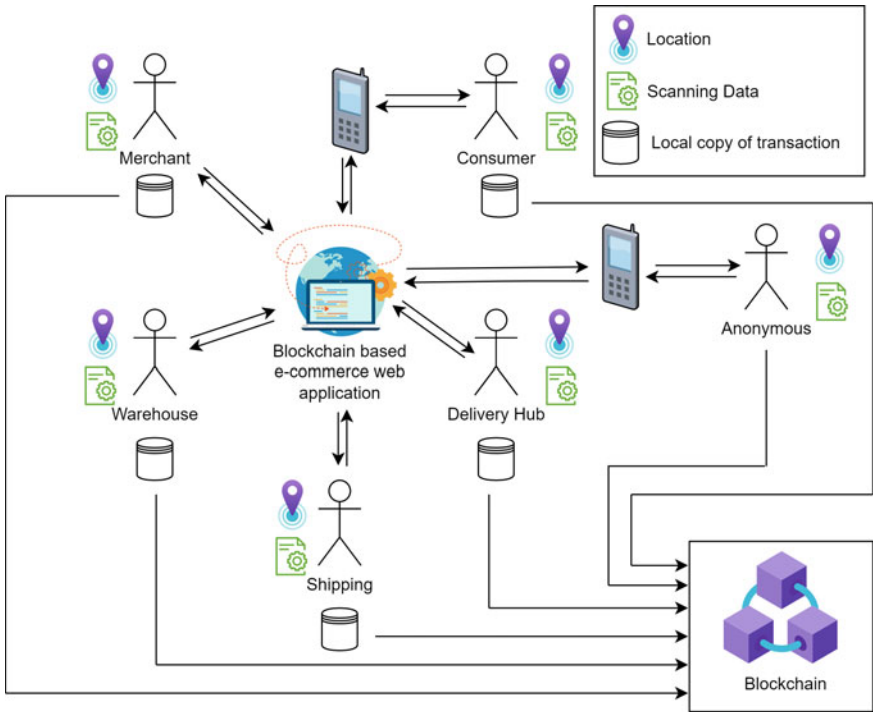


Fig. 3 Top-level model of the proposed system

a public and a private key and each user is linked with the unique product key. So there are few risks of identity theft. In the supply chain system, the users involved in transactions are known to each other and only known parties are allowed to participate. Here known parties mean merchant, warehouse, shipping, delivery hub, and consumer. Thus, a private blockchain is appropriate for our system model. At every stage when a user adds data, it is regarded as a transaction and the transaction data is written in the blockchain for data integrity, transparency, and accountability. Every user has his private key and public key in the system, and the public key is known to everyone. When a user adds data, the data is encrypted with the user's private key. This encrypted data and public key are hashed together to make a digital signature. A transaction contains a public key, digital signature, and the actual data, i.e., product-related information with location and device details. If there is an anonymous user, then the transaction contains only the actual data as there is no public–private key related to the anonymous user. This transaction is added to a block in the blockchain of the system. A block contains transaction data, a hash of the previous block, block creation timestamp, and a hash of the full block (Fig. 4).

Algorithm 2 Adding E-Commerce Transaction Data into Blockchain

Require: User's private key U_{pvk}, User's public key U_{pk}, Actual data A_d
Ensure: Transaction data T_d
function TRANSACTION (U_{pvk}, U_{pk}, A_d)
  if valid user **then**
    $E_d \leftarrow$ Encrypt (U_{pvk}, A_d)
    $H_d \leftarrow$ Hash (E_d, U_{pk})
    return $T_d \leftarrow U_{pk}, H_d, A_d$
  else
    return $T_d \leftarrow A_d$
  end if
end function

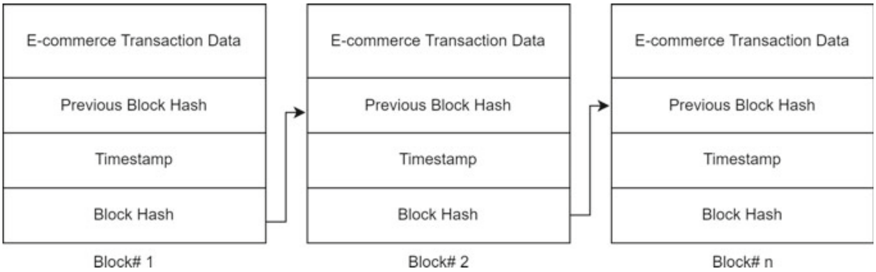


Fig. 4 E-commerce transaction data in the blockchain

4 Results

To conduct the experiment, we developed a web application and an android mobile application. PHP and MySQL were used to build the web application, and Java was used for the android application. There was an API between the two applications for communication. The web application was hosted on localhost using XAMPP software. All users except consumers used the web application, and the android mobile application was for the consumers.

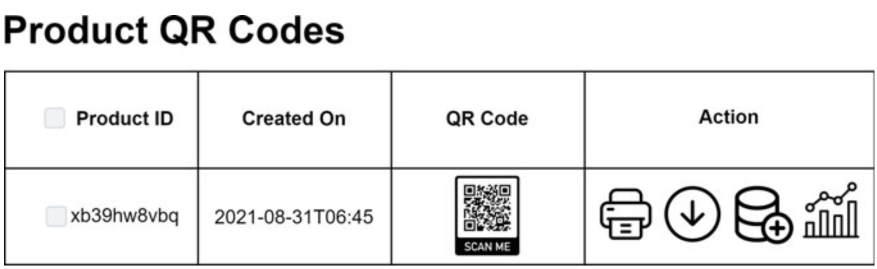


Fig. 5 QR code management by the warehouse in web application

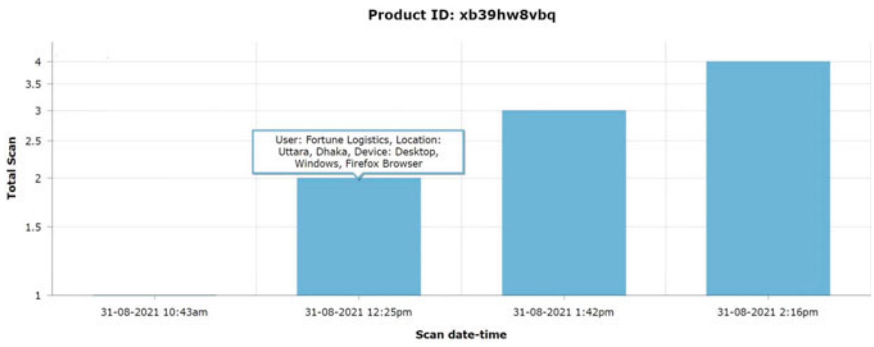


Fig. 6 Analytics of an individual product QR code

Figure 5 shows QR code options to the warehouse account. Here product QR codes are generated automatically according to the individual product when the consumers place orders. Even if a consumer orders for the same products, the product ID will be unique as is the QR code. As the QR code already contains product-related information, it saves data entry time and reduces errors. The warehouse has the options to print, download, add information, and view analytics related to QR codes. The other users have their related options according to their account type, i.e., merchant, shipping, etc. From the analytics, users can view the product-related information and the QR code scanning date-time, location, device types, scanned by whom, etc. These data can be used to track counterfeit products, and valuable business opportunities can be identified to make decisions and improve marketing return on investment.

Figure 6 shows some analytics of a scanned QR code. The figure is self-explanatory. It shows the scanning date-time, number of scanning, scanning location, scanned by whom, device information, etc. It is a historical data of the product. Every product having an unique ID has this individual information from where the product movement can be tracked. We took one hundred products as a dataset to conduct the experiment. Here only a single product information is presented as an example.

Figure 7 shows the steps of product verification by the consumer using the mobile application. The scanning result contains product-related information with a product journey map. By analyzing the scanning history, the app suggests in a popup window whether the product is original or counterfeit. This is done by fetching data from the blockchain by using the unique product ID. If the product-related users scan the QR code from logical locations and date-times with the correct sequence, then the product seems to be authentic. Here, the correct sequence means the product will be scanned by the shipping first when taking it from the warehouse, then by the consumer's nearest delivery hub, then by the shipping and finally by the consumer. If the product was scanned by an anonymous person or beyond related users, or the scanning sequences were random, or the scanning locations and date-times are unusual, then the product seems to be counterfeit. The consumer can reject and

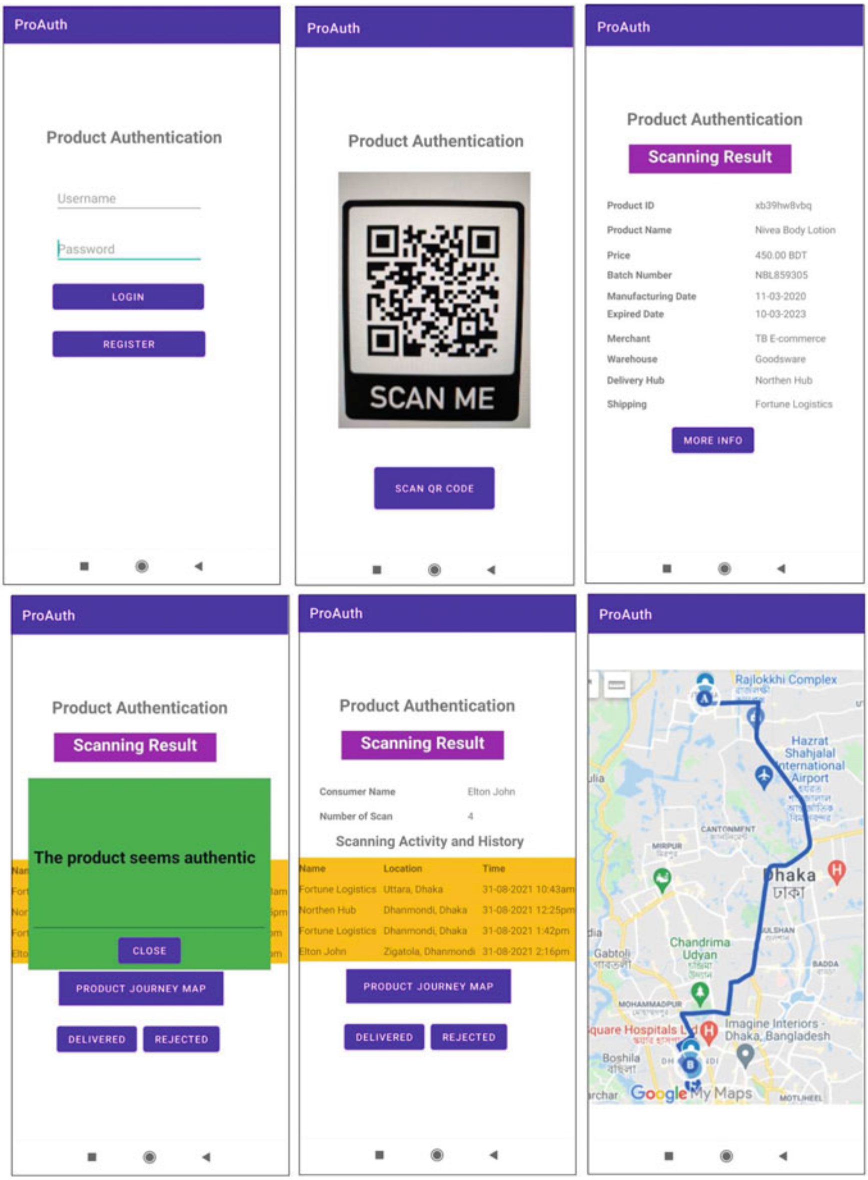


Fig. 7 Product verification by scanning QR code in mobile app

return the product with feedback in the app. Unusual scanning locations mean the product should be within a geographic area and should follow a normal path. If the warehouse location is 20 km away from the consumer, the product should not be 150 km away, and according to scanning locations, the distances should be lesser as date-time elapses.

Table 1 General e-commerce versus e-commerce with the proposed system

Context	General e-commerce	E-commerce with the proposed system
Product authentication	Limited and challenging	Improved and easier
Product mismatch	Frequent	Very less
Product tracking	Limited data	More data
Transparency	Limited visibility	Clear visibility
Trust among the users	Less	More
Identify malicious participant	No	Yes
Brand reputation	As usual	Increased
Revenue	As usual	Increased
Database tampering	Possible	Very hard
Health and safety safeguards	As usual	Increased
Environment friendly	Less	More
Online shopping experience	Average	Increased
Support for counterfeit	Very low	High
Data analytics	Less	More

Table 1 shows if the proposed system can be added to existing e-commerce, there will be a lot of improvements which were shown in the table context column that will reduce product counterfeiting and increase consumer satisfaction. An API can be developed to implement this system with the existing e-commerce websites so that there's no need to modify the existing e-commerce websites.

5 Conclusion

In this paper, the disadvantages of product counterfeiting and the relevant researches on product authentication systems have been discussed. We have shown that compared to the other systems which were discussed in the related works section, this system has much potential to reduce product counterfeiting as we have used blockchain along with destructible dynamic QR code and location tracking technologies to build the system. The web and mobile applications were built in such a way that data integrity, non-repudiation, and full traceability are ensured. This system is also user-friendly and has a very low implementation cost and complexity. Finally, a comparison is shown between general e-commerce and e-commerce with the proposed system where we met our goal to reduce product counterfeiting significantly.

In future work, machine learning and image processing algorithms on images of the product, brand logos, product description can be used to distinguish between counterfeit and original products, and data mining techniques can be used for analyzing sellers' reputations and consumers' reviews. A point-based scoring technique based on counterfeit product characteristics can be implemented to get product authenticity accuracy percentage in the scanning results. Besides, the Internet of Things (IoT) can be paired with the mentioned technologies for more features, accuracy, traceability, and transparency. Also, the proposed system can be extended for the users of root levels, i.e., raw materials suppliers to get authentic product ingredients information.

References

1. Dhanya, K.A.: Consumer protection in the e-commerce era. Int. J. Legal Res. (2015)
2. Otim, S., Grover, V.: E-commerce: a brand name's curse. Electron. Markets **20**, 147–160 (2010)
3. Raman, R., Pramod, D.: A strategic approach using governance, risk and compliance model to deal with online counterfeit market. J. Theor. Appl. Electron. Commer. Res. **12**, 13–26 (2017)
4. Bian, X., Veloutsou, C.: Consumers' attitudes regarding non-deceptive counterfeit brands in the UK and China. J. Brand Manag. **14**, 02 (2007)
5. Blackstone, E., Fuhr, J., Pociask, S.: The health and economic effects of counterfeit drugs. Am. Health Drug Benefits **7**, 216–224 (2014)
6. Cheng, M.: Is the drugstore safe? Counterfeit diabetes products on the shelves. J. Diabetes Sci. Technol. **3**, 1516–1520 (2009)
7. OECD and European Union Intellectual Property Office: Trends in Trade in Counterfeit and Pirated Goods (2019)
8. OECD and European Union Intellectual Property Office: Misuse of Small Parcels for Trade in Counterfeit Goods (2018)
9. Chatterjee, S., Datta, P.: Examining inefficiencies and consumer uncertainty in e-commerce. Commun. AIS **22**, 525–546 (2008)
10. Barkatullah, A.H., Djumadi: Does self-regulation provide legal protection and security to e-commerce consumers? Electron. Commer. Res. Appl. **30**, 94–101 (2018)
11. Xiao, B., Benbasat, I.: Product-related deception in e-commerce: a theoretical perspective. MIS Q. **35**, 169–195 (2011)
12. Cane, P., Primrose, D.: Food fraud: supply chain vulnerabilities and criminal opportunities during the covid pandemic. J. Appl. Anim. Ethics Res. **3**(1), 141–151 (2021)
13. Sotos, J., Houlding, D.: Blockchains for data sharing in clinical research: trust in a trustless world. Blockchain Application Note 1. Intel, Santa Clara, CA, USA (2017)
14. Sarmah, S.S.: Understanding blockchain technology. Comput. Sci. Eng. **8**(2), 23–29 (2018)
15. Trpovski, Z.: Geometric modifications of QR code. In: 2017 South Eastern European Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), pp. 1–6. IEEE (2017)
16. Yaga, D., Mell, P., Roby, N., Scarfone, K.: Blockchain technology overview, Oct 2018
17. Min, H.: Blockchain technology for enhancing supply chain resilience. Bus. Horiz. **62**(1), 35–45 (2019)
18. Sivaram, M., Laxmi Lydia, E., Pustokhina, I.V., Pustokhin, D.A., Elhoseny, M., Joshi, G.P., Shankar, K.: An optimal least square support vector machine based earnings prediction of blockchain financial products. IEEE Access **8**, 120321–120330 (2020)
19. Chen, W., Xu, Z., Shi, S., Zhao, Y., Zhao, J.: A survey of blockchain applications in different domains. In: Proceedings of the 2018 International Conference on Blockchain Technology and Application, pp. 17–21 (2018)
20. Ray, P.P., Dash, D., Salah, K., Kumar, N.: Blockchain for IoT-based healthcare: background, consensus, platforms, and use cases. IEEE Syst. J. **15**(1), 85–94 (2020)
21. Choi, S.H., Poon, C.H.: An RFID-based anti-counterfeiting system. IAENG Int. J. Comput. Sci. **35**(1) (2008)

22. Chen, Y., Chen, C.: Improve the performance of traceability system by using a digital certificate enabled anti-counterfeit QR-code mechanism. Int. J. Soc. Sci. Humanit. **7**(8) (2017)
23. Khalil, G., Doss, R., Chowdhury, M.: A novel RFID-based anti-counterfeiting scheme for retail environments. IEEE Access **8**, 47952–47962 (2020)
24. Li, Z., Liu, G., Liu, L., Lai, X., Xu, G.: IoT-based tracking and tracing platform for prepackaged food supply chain. Ind. Manag. Data Syst. (2017)
25. Haq, I., Esuka, O.M.: Blockchain technology in pharmaceutical industry to prevent counterfeit drugs. Int. J. Comput. Appl. **180**(25), 8–12 (2018)
26. Agrawal, T.K., Koehl, L., Campagne, C.: A secured tag for implementation of traceability in textile and clothing supply chain. Int. J. Adv. Manuf. Technol. **99**(9), 2563–2577 (2018)
27. Saeed, M.Q., Bilal, Z., Walter, C.D.: An NFC based consumer-level counterfeit detection framework. In: 2013 Eleventh Annual Conference on Privacy, Security and Trust, pp. 135–142. IEEE (2013)
28. Hu, S., Dai, T.: Online map application development using google maps API, SQL database, and ASP.NET. Int. J. Inf. Commun. Technol. Res. **3**(3) (2013)
29. Abraham, A., Santosh Kumar, M.B.: A study on using private-permissioned blockchain for securely sharing farmers data. In: 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pp. 103–106. IEEE (2020)
30. Banerjee, A.: Blockchain technology: supply chain insights from ERP. In: Advances in Computers, vol. 111, pp. 69–98. Elsevier, Amsterdam (2018)

Organizational Network Monitoring and Security Evaluation Using Next-Generation Firewall (NGFW)

Md. Shamimul Islam, Nayan Kumar Datta, Md. Imran Hossain Showrov, Md. Mahbub Alam, Md. Haidar Ali, and Md. Dulal Hossain

Abstract The most Bangladeshi government and non-government scientific organizations have comprised both wired and wireless system network infrastructure. When such network infrastructure has been designed and developed at these organizations, ensuring network security has become a highly challenging issue. For this purpose, Next-Generation Hardware Firewall (NGFW) has been implemented by most scientific organizations. This hardware firewall is a very effective technology to perform network security. In this paper, we have chosen Atomic Energy Research Establishment (AERE), Savar under Bangladesh Atomic Energy Commission (BAEC) as an experimental zone. To facilitate research and services, Internet-based network infrastructure has been established at AERE, BAEC. A Next-Generation CISCO-based network firewall has been integrated with the implementation of AERE network infrastructure to enhance and ensure network security. We have experimented and collected data from October 09, 2020 to October 09, 2021 at the network system of AERE. By collecting and evaluating these statistical data in terms of different network security parameters, we have investigated AERE's network data like controlling ingress and egress traffic, monitoring traffic flow, malware threat analysis, web content and application filtering, URL filtering, decrypting traffic results, etc. In this paper, we have illustrated how network and application security is being achieved

Md. S. Islam (✉) · N. K. Datta · Md. I. H. Showrov · Md. M. Alam · Md. H. Ali · Md. D. Hossain
Institute of Computer Science, Bangladesh Atomic Energy Commission, Agargaon, Bangladesh
e-mail: shamimul.jucse@gmail.com

N. K. Datta
e-mail: nayan.baec@gmail.com

Md. I. H. Showrov
e-mail: showrov.cse@gmail.com

Md. M. Alam
e-mail: mahbubcseju@gmail.com

Md. H. Ali
e-mail: haiderdiu@gmail.com

Md. D. Hossain
e-mail: edhossain@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_10

by the NGFW firewall and as a resulting factor, the summarization of scale up the network security for a scientific organization is enlightened.

Keywords Firewall · Network security · Networks traffic · Threats · Intrusion

1 Introduction

Bangladesh has a goal of digital Bangladesh-2021. For this purpose, with the rapid advancement of ICT in the world, Bangladesh is also moving into the digitalization process fast in every sector like government organizations, offices, educational institutions, etc. For this reason, every institution is establishing a network-based infrastructure (Internet, Intranet, Extranet, etc.). From this perspective, maintaining network security becomes very important. Because insufficient network security causes irreparable damages which indeed hinders creating our digital Bangladesh of vision-2021, which is not desirable. Due to improper network, security implication causes potential threats. It creates confidentiality violations especially for maintaining organization web form data. So there is a need to use a strict network policy to ensure prevents potential losses regardless of size and type of network. Network policy means a strong set of rules, regulations, and structured designed network framework developed and implied by the organizational administrators to prevent and monitor unauthorized access, misuse of network, correction, and restriction, ensure confidentiality and integrity for the network-accessible resources [1]. To ensure network security, firewall strategies are a much effective process for ensuring organizations' security policy. It can protect a network from various threats in terms of different network security parameters. It generally protects computer and communication technology instruments within the network framework. Security is the main purpose of using a firewall, as we have to determine a key to how much protection is needed.

Firewalls can be of two types: hardware and software. A software firewall is generally ingrates with a specific host Operating System (OS). Software firewall generally plays a role internally in a host computer through applications and port numbers. Its implication is limited. In this paper, we have discussed hardware firewalls for much bigger network and communication technology security. A hardware firewall physically exists which is stored between a network and gateway. Different firewall performs different roles from a different perspective. In organizations, network infrastructure hardware firewalls are being implemented, and it monitors or filters ingress or egress traffic basically in transport layer 4 on the OSI 7 layers model. High-level programming languages are being integrated to define simplify firewall policy. On the other hand, the Next-Generation Firewall can be operated on all layers of the OSI 7 layer model.

In this paper, we have analyzed the experimental ingress and egress data of the network using Next-Generation Firewall at Atomic Energy Research Establishment (AERE), Savar of Bangladesh Atomic Energy Commission. This paper provides information on how Next-Generation Firewall (NGFW) policy has been set to protect the AERE's network, and network security is being achieved in terms of identifying and controlling applications, virtualization of software firewall, managing traffic especially unknown traffic, controlling user, intrusion prevention and detection against known threats, prevents unknown threats, track capabilities of users, applications, location via geo-location database, ports and session, and IP addresses.

The organization of the remaining paper is as follows. Section 2 describes a detailed literature review of recent related works. The methodology or designed network and its security framework of the research have been illustrated in Sect. 3. Data analysis and obtained results are analyzed in Sect. 4 along with some recommendations. Finally, the paper is concluded in Sect. 5 along with the future directions.

2 Related Works

Most of the current researches on network security and Next-Generation Firewall are concerned about how network security is violated in different areas, configuration and misconfiguration of firewall and performance analysis of different firewall in different criteria. However, very few papers have discussed designing firewall framework policy and security evaluation according to specified designed framework policy in terms of different network security parameters. The research [2] provides researchers with a current overview of the modern network monitoring approach. In this activity, the writer deals with high-speed bandwidth Internet in both wired and wireless networks.

In [3], the researcher did a literature review on firewall security issues various policies and the concept of distributed firewall. A Review in Recent Development of Network Threats and Security Measures is discussed at [4]. Their research was mainly concerned with increased security-related studies with the web data. Hayajneh et al. [5] examined various types of firewalls operations. They tested the performance and security for various firewalls including CISCO ASA, packet filter, and Checkpoint SPLAT. The researchers also studied the effect of implementing a firewall on network performance. Konikiewicz and Markowski [6] have worked on the performance analysis of CISCO ASA and Juniper firewalls on packet traffic based on bandwidth and server response time.

In [7] and [8], the authors presented a paper on the firewall performance of firewall using the Markov chain. The methodology analyzes firewalls that are subject to normal traffic flows as well as denial of services (DoS) attack flows. Other literature [9] and [10] presented the idea to define the role of firewall on computer networks and implementation of firewall on hardware or software on combination both. Zaki et al. [11] conducted a survey on various advanced attacks and damages that occurred on the healthcare system using the existing models. Finally, they proposed an FGFW-based model and configured it with a distinctive hospital network. Neupane et al. [12] presented a survey on different firewalls highlighting network security. In addition, they illustrated the advantages were being visualized including the primary goals of network security.

All mentioned research works show how the firewall is implemented, configuration, why needed, performance analysis of different firewall by category the firewall implementation to analysis into a smaller subcategory. A real-time data analysis by implementing hardware firewalls is being avoided in these mentioned research works. However, we are unable to get more statistical data for a scientific organization as the firewall is working as a network security instrument.

3 Design Network and Security Framework

To design a secure network infrastructure, we have chosen the AERE, Savar of BAEC as a model experimental area. AERE has been developed a huge network both comprised of wired and wireless systems. A significant number of LAN connections including switches, routers, and media converters exist to broadcast data of its remotely located institutes. In addition, ingress traffic and egress traffic have been controlled by implemented NGFW firewall. The ingress traffic for the different host within the LAN network and Demilitarized Zone (DMZ) servers is being monitored by this NGFW firewall from the Internet and outside servers. DMZ network is used to create a new layer that protects the scientific organizations' LAN from unauthentic access. In DMZ zone, basic resources are DNS, FTP, proxy, mail, web server, and Voice over Internet Protocol (VoIP), etc. (Fig. 1).

This firewall is used to define a set of rules which enables the firewall to check the data packets coming from the Internet, extranet, servers, or any external networking system to the LAN if there are no obstacles found and vice versa. The firewall drops the packet, and an error is detected by detecting harmful vulnerability in the data packets coming from the PC. The firewall authorizes the ingress access and egress data according to its defined rules.

The set of rules or policies can be applied by a firewall. The NGFW firewall features perform their tasks according to these set rules for multilayer protocols in the OSI model. The NGFW firewall needs to set policy and rules in a specific way to determine which activity should be performed.

For segmentation, NGFW creates zones that prevent unauthorized connections and make successful connections. Such segmentation examples are clusters of inside zone IPs to WAN and vice versa which is shown in Figs. 2 and 3.

NGFW can be used as both intrusion prevention system (IPS) and intrusion detection system (IDS), whereas IPS and IDS systems have their own defined signatures, algorithms, and heuristics to detect and block intrusion or attack on the network and host [13]. To configure Intrusion Prevention System (IPS), every NGFW firewal has its extended sensor which is deployed with the service router. By implementing policies on NGFW firewall, the host sensors is able to communicate with the service router. For the inspection of egress or ingress traffic for LAN-to-WAN or WAN-to-LAN, the sensor of the NGFW firewall activates its sensors to perform IPS.

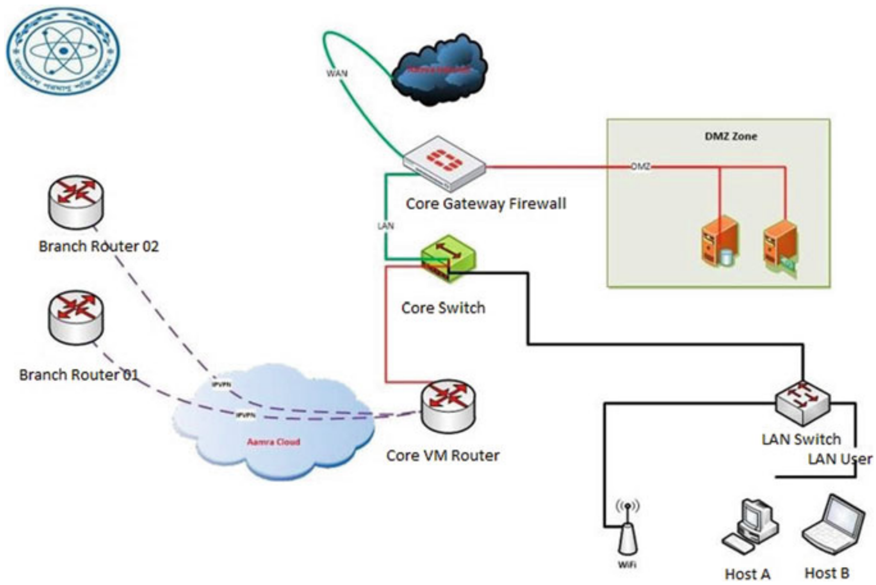


Fig. 1 Designed network of AERE, BAEC

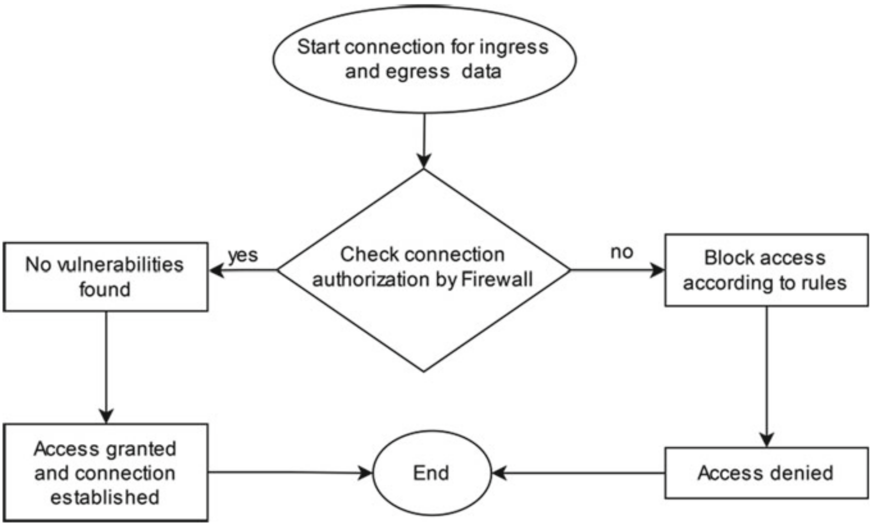


Fig. 2 Traffic access flow using NGFW firewall

Fig. 3 Segmentation using NGFW firewall

Content and application filtering are being done by the NGFW firewall in a very systematic and modern way. For URL filtering, rules are being applied in the NGFW firewall by the authority. Only acceptable categories of domains are allowed in this process. Categories within a domain may include different websites which are related to hacking, nudity, phishing or violence, etc. Similarly, specific sessions, ports and services, and IP addresses can be ruled for filtering to make acceptable categories. For controlling application, the NGFW firewall is not only applicable in layer 4 but also control applications in layer 7 (application layer) of the OSI model. It decodes network streams and determines presented content in the application layer. In Figs. 4 and 5, different policy implementation like URL filtering and application is shown. Similarly, port scanning and individuals can be filtered out by implementing policy.

The NGFW can control contents and act as a sandboxing system. In terms of content control, NGFW firewall tries to reveal specific content within the application such as word documents, pdf files, executive files, and scripts. The NGFW firewall applies a user-defined malware policy or default policy to detect malware of files before being downloaded or uploaded. It detects host receiving malware and malware

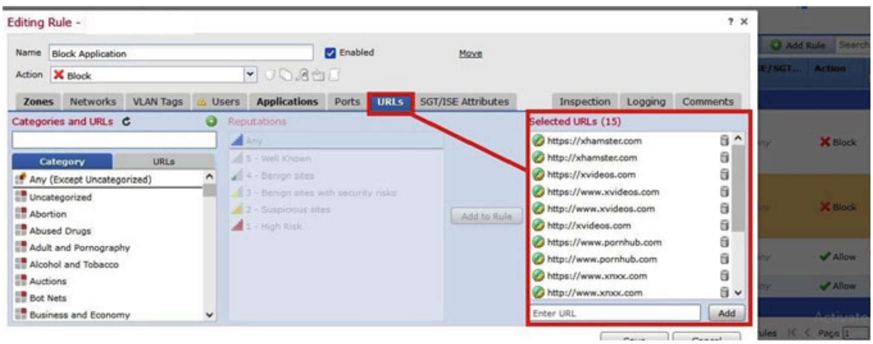


Fig. 4 URL policy add in NGFW firewall

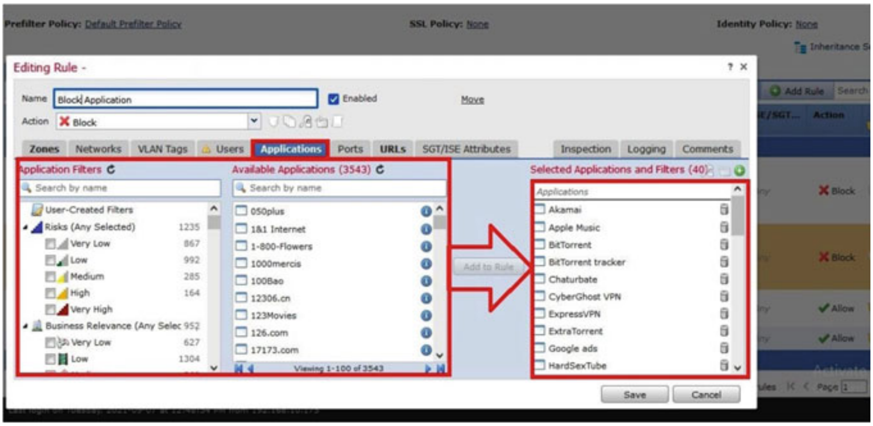


Fig. 5 Application filtering add in NGFW firewall

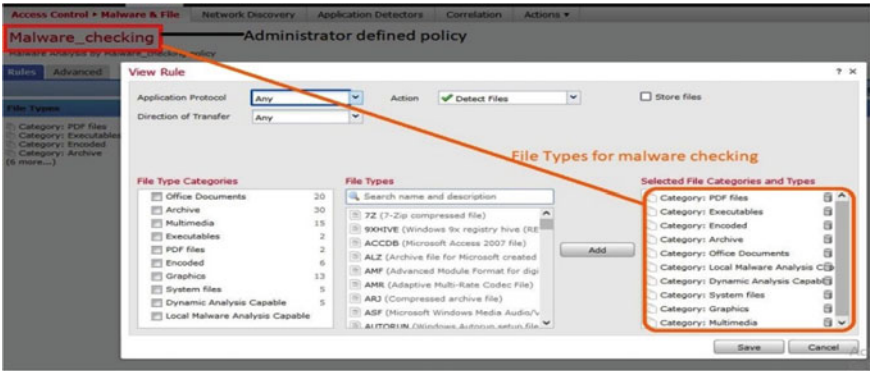


Fig. 6 Malware policy using integrated sandboxing in NGFW firewall

intrusion of a network trajectory based upon applied policy. In Fig. 6, malware policy using CISCO integrated Sandboxing in the NGFW firewall is shown. The NGFW also acts as web application firewall (WAF) which adds more features in HTTP protocol to detect and stop vulnerable network threats.

4 Experiment Data and Result Analysis

This section provides a detailed description of obtained results of the NGFW firewall at the AERE perspective from October 09, 2020 to October 09, 2021.

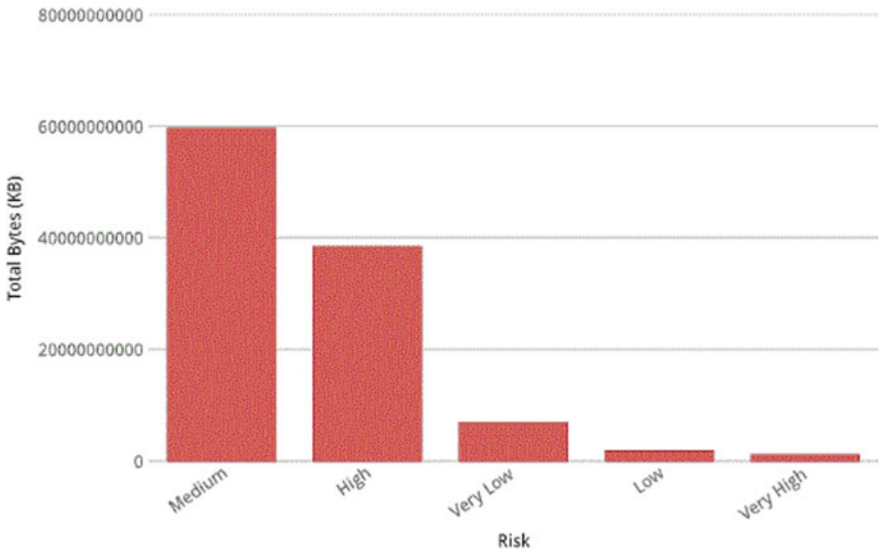


Fig. 7 Traffic by application risk by initiator IP at AERE

4.1 Network Traffic or IP Monitoring

To track active network connections, the NGFW plays a crucial role to present data in the application form in the presentation layer by decoding network streams. In Fig. 7, we can see the data traffic by application risk of the end user of AERE internal LAN zones based on predefined firewall policy, whereas Fig. 8 shows which category application consumes most traffic within the network.

Based upon the results, the NGFW gives us a clear view of how to monitor the traffic of AERE LAN and bandwidth broadcast view. It helps to make a decision which internal applications and users are acting as most suspicious and should be monitored and a cluster of users should be done based on user traffic and application.

4.2 On the Perspective of Intrusion Detection and Prevention

From the establishment of NGFW in AERE, we have found that it is performing in an excellent way to detect network Trojan-type intrusion and drop or block that threat event. It has protected AERE's internal network by allowing non-malicious traffic from the Internet and other unknown network zones.

In Fig. 9, we observed that in AERE's network, we get the data traffic and intrusion events over one year time period, and in Fig. 10, we got the result of high priority worse network Trojan intrusion events, or malware was detected and dropped by NGFW.

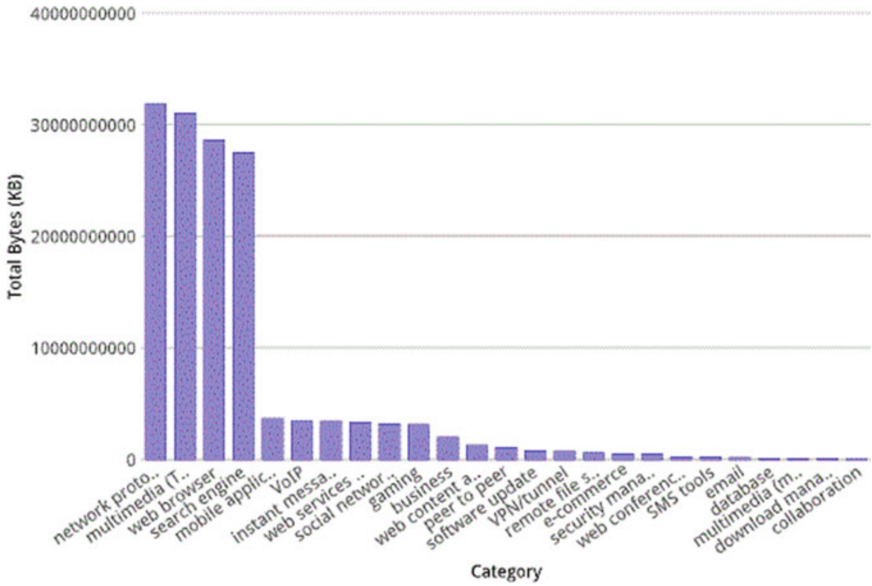


Fig. 8 Consume data traffic by application category at AERE

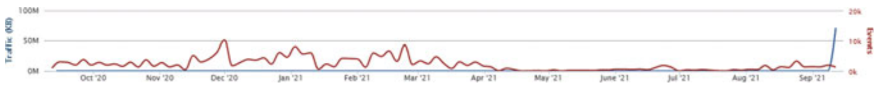


Fig. 9 Traffic and intrusion events over time

Event	Classification	Priority	Events
MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35549:1)	A Network Trojan was Detected	high	50210
MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)	A Network Trojan was Detected	high	49447
MALWARE-CNC Win.Trojan.Glupteba.M initial outbound connection (1:30288:2)	A Network Trojan was Detected	high	31421
MALWARE-CNC Win.Trojan.Nvbpass variant outbound connection (1:19864:6)	A Network Trojan was Detected	high	31367
MALWARE-CNC DNS suspicious .bit dns query (1:41083:4)	A Network Trojan was Detected	high	21960
MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35746:1)	A Network Trojan was Detected	high	17000
MALWARE-CNC Win.Worm.Brontok user-agent outbound connection (1:20021:7)	A Network Trojan was Detected	high	2611
PUA-ADWARE Slimware Utilities variant outbound connection (1:46486:1)	A Network Trojan was Detected	high	227
EXPLOIT-KIT CritX exploit kit Portable Executable download (1:24791:4)	A Network Trojan was Detected	high	196

Fig. 10 Intrusion events details detected by NGFW

4.3 On the Perspective of Content and File Filtering and Disposal of Malware

The NGFW of AERE performs at a satisfactory level in terms of URL filtering, content filtering, and application monitoring with the protocols also.

In Fig. 11, we can see that the NGFW firewall monitored ZIP, Pdf, Word, mp3, mp4, mscab, and other relevant types of files. It identified malware, proprietary, and confidential information and disposed malware from the network. In Fig. 12, we have got the investigated result of the suspicious malware to these files and disposal of malware from these files within the time frame.

For application control, the NGFW gives us the data of AERE's network in terms of allowed or denied connection by application, allowed connection by application risk and business relevance, traffic consumed by application and application category, etc.

In Fig. 13 and Table 1, we are able to gather accurate information about the usage of applications based on traffic, risk category, allowed or accessed application, and denied application statistics according to set up NGFW policy at AERE.

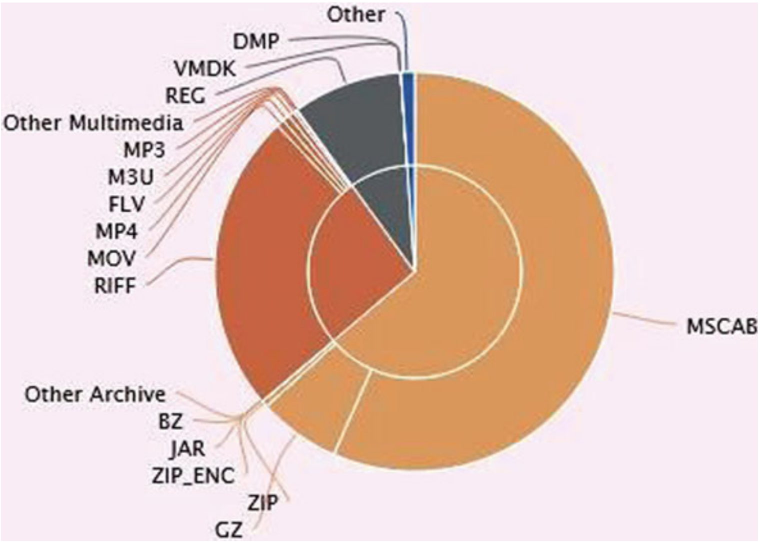


Fig. 11 Files types are being monitored and downloaded within AERE

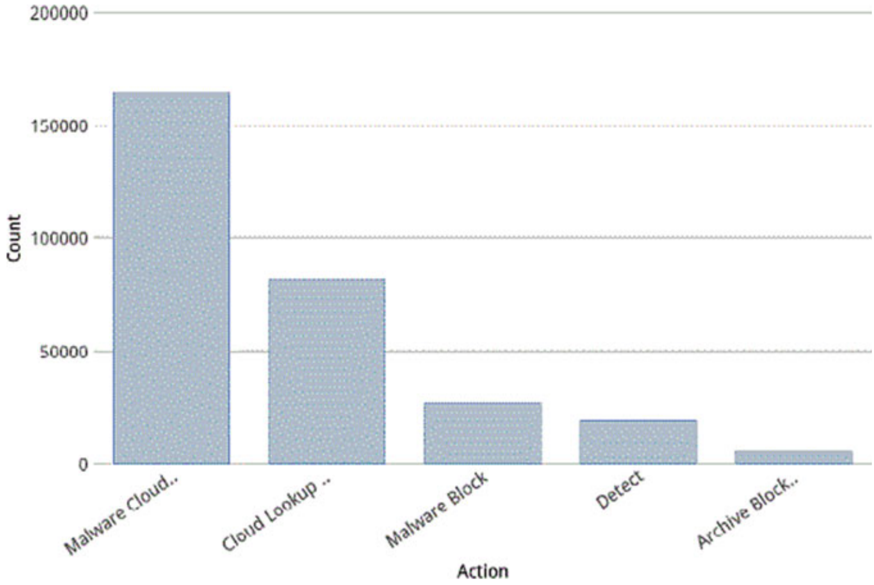


Fig. 12 Type of actions on files in AERE network

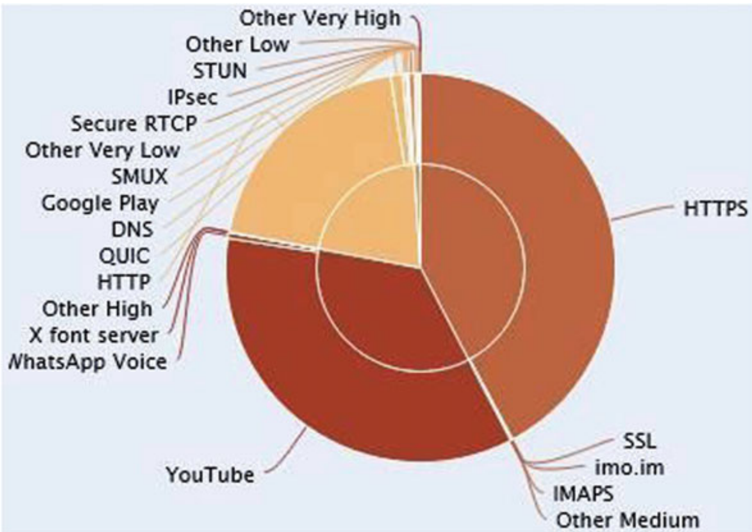


Fig. 13 Usage of traffic by risk and application

Table 1 Number of connections by application risk policy based on NGFW

Application	Allowed connection	Denied connection
ICMP	172,154,365	–
ICMP client	172,154,365	–
DNS	122,989,129	–
DNS client	122,859,033	–
HTTPS	83,634,938	756
YouTube	23,245,865	–
BitTorrent	14,101,160	–
Facebook	10,012,907	21
HTTP/HTTPS	9,798,159	43,355
Google	5,933,624	–
XVideos	–	38,970
Chaturbate	–	1536
Chrome	933,624	221

4.4 *NGFW Sandboxing and Malware and Threat Analysis Based on Attacks*

Like every sandboxing facility process, AERE established NGFW has contained an integrated cloud-based sandboxing facility which is very much effective to detect and drop different threats and malware against denial of services attack, networks worms and Trojan detection, ransomware, spoofing and unauthorization, key logging, etc. In this section, we have found the results that the application protocol introduces about 28,281 malware during the last year. Table 2 gives us detailed data about client applications and web applications that introduces malware for one year.

From Fig. 14, we found the different malware threats over different files by their count. Malware content can be attached within any files when being downloaded

Table 2 No. of client/web application list introducing malware

Client/web application	No. of malware detection
Web browser	28,206
BitTorrent	32
Chrome	12
Android browser	8
Internet Explorer	6
uTorrent	32
Amazon Web Services	1
Facebook	–
Google	43

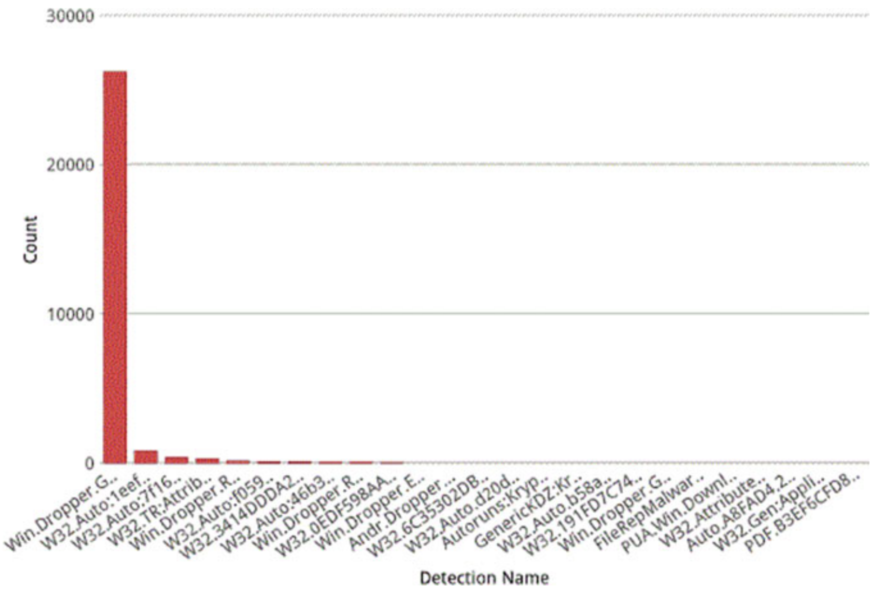


Fig. 14 Detected malware threats by NGFW

or uploaded or accessed on any websites or with any suspicious mail attachment. It creates a vulnerable situation to a network and associated host within a network.

NGFW firewall detects specific malware content or ransomware and takes proper action like malware block, drops, the cloud looked up which are given in Fig. 15. On the other hand, Fig. 16 gives us a clear overview that identified malware contents on different files to protect networks and files from different attacks using port scanning at AERE.

The threats that are entering a network can be prevented by a common technique called port scanning by the NGFW firewall [14]. Port scanning can be done to determine which service we can connect to. AERE's NGFW is quite well to detect malware-type suspicious activity by port scanning.

5 Conclusion

To protect a network system against different internal and external threats, improvement of network security is an ongoing and challenging issue for any scientific organization. As a scientific organization, we always need to analyze the network data and take proper steps to improve its network and application security area. In this research, we have analyzed the various challenging and evolving network issues of modern-day's technology. Considering security parameters, we have configured a CISCO-based NGFW hardware firewall over the network framework of AERE.

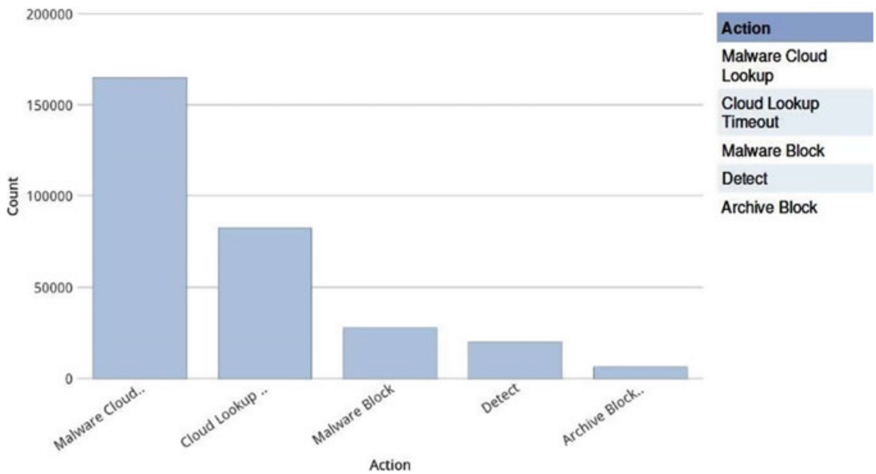


Fig. 15 Taken action against malware by NGFW

Source Port / ICMP Type	Destination Port / ICMP Code	SSL Status	VLAN ID	Message
49316 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
50513 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
49534 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
49531 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
49526 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
1250 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
50508 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
49523 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)
49522 / tcp	80 (http) / tcp	Unknown (Unknown)	0	MALWARE-CNC Win.Trojan.Zeus variant outbound connection (1:35030:1)

Fig. 16 Detection malware message by port scanning using NGFW

Then, we have set different firewall policies and rules according to AERE network security perspectives and tried to visualize how this NGFW firewall protects our defined network. Our study investigated that NGFW plays a significant role to protect our network against unauthorization, denial of services attack, spoofing, unauthorized traffic, block malware content files and suspicious application, intrusions and network activity, etc. In the future, we will try to find out the better approach of VPN integration for remote access, security of the different embedded systems, and advanced web systems assimilation on the NGFW firewall to ensure network security at a large scale. We will also try to build a multilayer security system with NGFW and performance analysis compared with the existing system.

Acknowledgements The authors would like to express their gratitude to Nuclear Cyber Security LAB of Institute of Computer Science, BAEC, to provide data and conduct this research work.

References

1. Alsaqour, R., Motmi, A., Abdelhaq, M.: A systematic study of network firewall and its implementation. IJCSNS Int. J. Comput. Sci. Netw. Secur. **21**(4) (2021)
2. Svoboda, J., Ghafir, I., Prenosil, V.: Network monitoring approaches: an overview. Int. J. Adv. Comput. Netw. Secur. **5**(2), 88–93 (2015)
3. Chopra, A.: Security issues of firewall. Int. J. P2P Netw. Trends Technol. (IJPTT) **22**(1), 4–9 (2016)
4. Dastres, R., Soori, M.: A review in recent development of network threats and security measures. Int. J. Inf. Sci. Comput. Eng. (2021)
5. Hayajneh, T., Mohd, B.J., Itradat, A., Quttoum, A.N.: Performance and information security evaluation with firewalls. Int. J. Secur. Appl. **7**(6), 355–372 (2013)
6. Konikiewicz, W., Markowski, M.: Analysis of performance and efficiency of hardware and software firewalls. J. Appl. Comput. Sci. Methods **9** (2017)
7. Salah, K., Elbadawi, K., Boutaba, R.: Performance modeling and analysis of network firewalls. IEEE Trans. Netw. Serv. Manage. **9**(1), 12–21 (2011)
8. Trabelsi, Z., Zeidan, S., Hayawi, K.: Denial of firewalling attacks (DOF): the case study of the emerging blacknurse attack. IEEE Access **9**(7), 61596–61609 (2019)
9. Dandamudi, S., Eltaeib, T.: Firewalls implementation in computer networks and their role in network security. J. Multi. Eng. Sci. Technol. (JMEST) **2**(3) (2015)
10. Islam, S., Ali, H., Habib, A., Nobi, N., Alam, M., Hossain, D.: Threat minimization by design and deployment of secured networking model. Int. J. Electron. Inf. Eng. **8**(2), 135–144 (2018)
11. Zaki, M., Sivakumar, V., Shrivastava, S., Gaurav, K.: Cybersecurity framework for healthcare industry using NGFW. In: Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 196–200 (2021)
12. Neupane, K., Haddad, R., Chen, L.: Next generation firewall for network security: a survey. SoutheastCon **2018**, 1–6 (2018)
13. Cyber Security Tutorial Homepage. https://www.w3schools.com/cybersecurity. Last accessed 14 Sept 2021
14. Port Scanning. www.barracuda.com/glossary/port-scanning. Last accessed 29 Nov 2021

Crypto Holographic Reconfigurable Intelligent Surface-Assisted Backscatter Communication

Md. Saiam, Mostafa Zaman Chowdhury, and Syed Rakib Hasan

Abstract The upcoming sixth-generation (6G) mobile network introduces different communication technologies for efficient and controllable propagation. Backscatter communication (BackCom) is one of those which can transmit signals using the passive tag. This BackCom technology can harvest energy from the ambient signal sources (such as television and mobile phone tower) and is very energy efficient for 6G network and Internet of Things communication. For smooth and controllable propagation of signals, the reconfigurable intelligent surface (RIS) communication technology is very effective. With the help of its electronically controlled reflected path, the RIS technology improves the signal and network quality for better performance. This paper presents a crypto holographic RIS-assisted BackCom for secured and lossless communication in the upcoming network technologies. This combined communication technology will be energy efficient. It will increase the signal strength and data rate which are very important for the next generation mobile network. Research findings clearly show that when we use RIS-assisted BackCom, the channel gain, data rate as well as energy efficiency of signal reflecting surface are also increased significantly.

Keywords Backscatter communication · Internet of Things · Reconfigurable intelligent surface · Reflector

1 Introduction

With the growth of communication sector, the efficiency and performance of wireless networks have improved a lot. From the first-generation to sixth-generation

Md. Saiam (✉) · M. Z. Chowdhury · S. R. Hasan
Khulna University of Engineering and Technology (KUET), Khulna 9203, Bangladesh
e-mail: mdsaiam.eee@gmail.com

M. Z. Chowdhury
e-mail: mzaman@eee.kuet.ac.bd

S. R. Hasan
e-mail: syedrakibhasanadit@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_11

(6G), mobile networks have done mass changes in wireless communication. The 6G mobile network will be a single platform for providing multiple types of services and applications [1]. Day by day, the demand for data is increasing with the development of massive communication of multiple-input multiple-output (MIMO) devices, device-to-device (D2D) communication, Internet of Things (IoT), and smart data communication. The main objectives of the 6G network are to improve data rate, capacity, a better quality of services (QoS), and lower latency with high speed, and lossless data communication.

The controllable propagation of the electromagnetic wave (EM) is one of the challenges for the next generation mobile network [2]. The researchers have proposed the reconfigurable intelligent surface (RIS) and observed it for the application in wireless communications [3]. An RIS plate consists of many unit cells that can change the amplitude, polarization, and phase of incident signal with the help of the smart controller of RIS. So, the unit cells of RIS are capable of controlling the scattering of EM signals [4].

In recent times, backscatter communication (BackCom) is one of the promising communication systems for low power energy consumption which can be produced from the ambient signal. The wireless tag devices of the BackCom system modulate, reflect the signals, and also able to harvest energy from the signals [5]. A lot of IoT devices will be embedded in surroundings to support D2D communication. However, it will be very big challenge to support energy for these massive number of IoT devices. These IoT devices require low power which can be harvested using BackCom from ambient signals such as radio, television (TV), and mobile network [6]. To make smooth, strong, and efficient communication, researchers focus on combining communication technology and working on the RIS-assisted BackCom. The RIS can enhance the signal strength and performance using an electronically controlled reflected surface [7]. The BackCom cannot control and improve the propagation of EM waves and signal gain, respectively, which supports lower coverage. On the contrary, RIS system can improve signal gain and control EM propagation. To run the RIS system, it requires an additional energy source where the BackCom does not require any as it can harvest energy from the ambient sources.

In this paper, we proposed a RIS-assisted BackCom technology for better performance, secured, and energy efficient communication. Here, we use a crypto holographic tag for data encryption and secured communication. The data rate of this combined communication technology increases with the improvement of the probability of channel gain.

The rest of the paper is arranged as follows. In Sect. 2, we specify the system overview of crypto holographic RIS-assisted BackCom. Section 3 presents several derivations and comparisons of different systems. Performance analyses are discussed in Sect. 4, and conclusions are shown in Sect. 5.

2 System Overview

The proposed architecture of the RIS surface is formed with two-dimensional synthetic electromagnetic materials, which is known as metasurface. This metasurface is a form of several-number of arrays that consists of a huge number of reflecting components called meta-atom [8]. The signal response including amplitude and phase shift of reflected signal of each metasurface can be individually controlled by geometrical shape, architecture, and designing elements. The reflection parameter of each metasurface should be designed in such a way that can be tunable to support the random mobility of wireless networks. This property can be achieved by using electronic devices or micro-electromechanical actuators. A crypto holographic tag is integrated with the surface to encrypt the data.

In our proposed model as shown in Fig. 1, the system transmits the signal in five different ways the signal keeps encrypted using the crypto holographic technique. The signal sources are integrated with the cryptography which encrypts data before transmission. It integrates the crypto holographic tag in both RIS and BackCom tags so that if the signal comes from any ambient sources which are not encrypted such as mobile and TV networks, then the integrated crypto holographic tag of the nearest RIS or BackCom surface converts the data into the encrypted ciphertext for secured data transmission. After getting the encrypted ciphertext, the receiver decrypts it using the decryption algorithm which converts into the original data signal. This model enhances the signal quality and transmission range as well as reduces the signal losses. Due to low power consumption of backscatter surfaces, this crypto holographic RIS-assisted BackCom system can be a promising and effective communication medium for upcoming 6G network.

3 System Analysis

Figure 2 shows different transmission links of our proposed systems. Let $h_{\mathrm{sb},k}$, h_{sr}, $h_{\mathrm{sc},k}$, $h_{\mathrm{br},k}$, $h_{\mathrm{o},k}$, and $h_{\mathrm{cr},k}$ indicate the channel coefficient from signal source to backscatter tag k, from signal source to receiver, from source to RIS assisted backscatter reflector k, from backscatter tag k to receiver, between backscatter tag k, and RIS reflector k of RIS-assisted backscatter system, and from RIS-assisted backscatter reflector k to receiver, respectively.

Again, d_{sb}, d_{sr}, d_{sc}, d_{br}, d_{o}, and d_{cr} denote the distance from signal source to backscatter tag k, from signal source to receiver, from signal source to RIS-assisted backscatter reflector k, backscatter tag k to receiver, between backscatter tag k, and RIS reflector k of RIS-assisted backscatter system and from RIS-assisted backscatter system to receiver, respectively. Mathematically, the received signal by the kth backscatter surface, denoted by $y_{p,k}(n)$ and can be expressed as

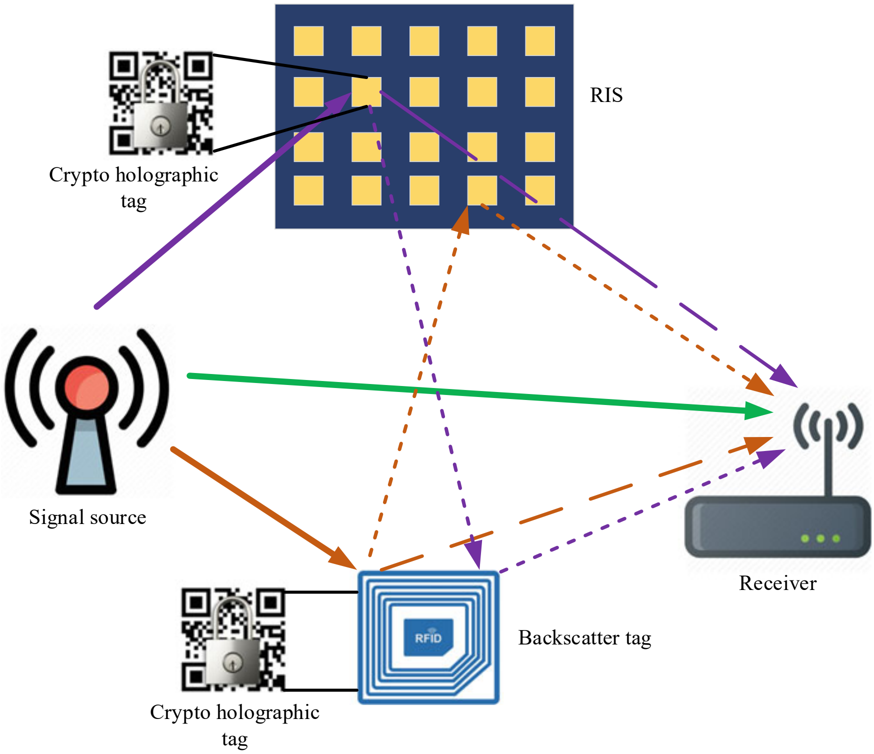


Fig. 1 System model

$$y_{p,k}(n) = \frac{h_{\mathrm{sb},k}}{\sqrt{d_{\mathrm{sb}}^{\alpha}}} s(n) \tag{1}$$

The received signal by the kth RIS-assisted backscatter system, denoted by $y_{q,k}(n)$ and can be expressed as

$$y_{q,k}(n) = \frac{h_{\mathrm{sc},k}}{\sqrt{d_{\mathrm{sc}}^{\alpha}}} s(n) \tag{2}$$

where α is large scale link loss factor for each channel and $s(n)$ is the transmitted data signal from the signal source.

The signal reflected by kth backscatter tag, denoted by $y_{b,k}(n)$ and can be expressed as

$$y_{b,k}(n) = y_{p,k}(n)x(n)e^{j\theta_k} \tag{3}$$

where $x(n)$ denotes the data.

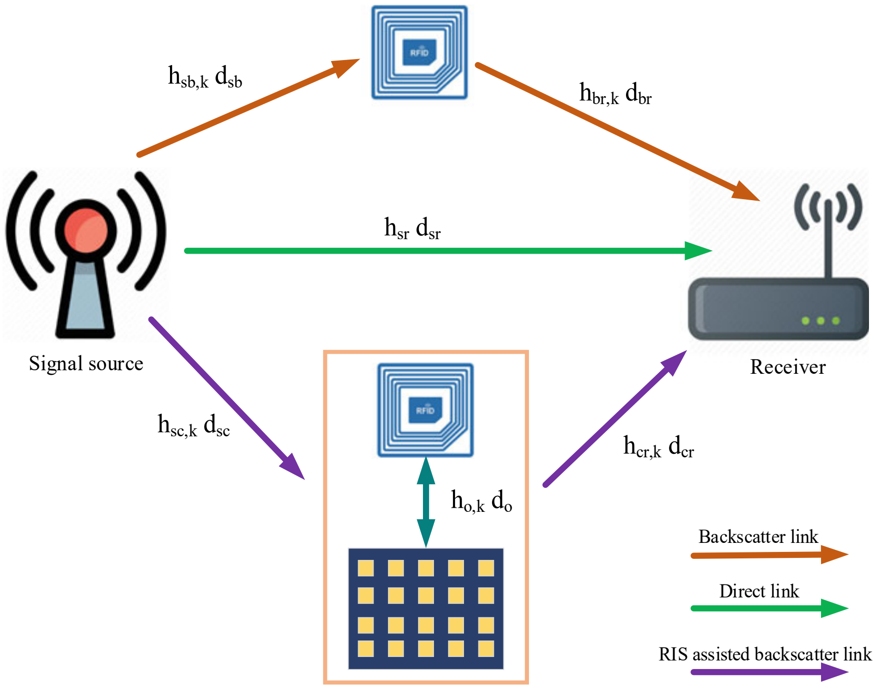


Fig. 2 Different transmission links

The signal reflected by kth RIS-assisted backscatter reflector, denoted by $y_{r,k}(n)$ and can be expressed by

$$y_{r,k}(n) = \eta y_{q,k}(n)x(n)e^{j\theta_k} \tag{4}$$

where η is the attenuation factor for RIS system.

The signal appeared at the receiver is the superposition of all the reflected signals and direct transmitted signal, denoted by $y(n)$ and expressed as

$$y(n) = \frac{h_{\mathrm{sr}}}{\sqrt{d_{\mathrm{sr}}^{\alpha}}}s(n) + \sum_{k=1}^{K} y_{b,k}(n) + \sum_{k=1}^{K} y_{r,k}(n) \tag{5}$$

Channel gain probability of the RIS-assisted BackCom link is bigger than that of the straight or only backscatter link [9]. Mathematically, which can be expressed as

$$P = P_{\mathrm{r}}(|G_1|\langle|\eta G_3|) \tag{6}$$

$$P = P_{\mathrm{r}}(|G_2|\langle|\eta G_3|) \tag{7}$$

where P_{r} is the received power of reflected signal and K is the total number of reflectors. G_1, G_2, and G_3 are the channel gain of direct communication link, BackCom link, and RIS-assisted BackCom link, respectively.

Equations (6) and (7) show the comparison of channel gain probability of RIS-assisted BackCom with direct communication and only backscatter communication. The channel gain probability of RIS-assisted BackCom is greater than direct or only backscatter communication. By using the proper combination and number of reflectors, it can be possible to get better channel gain.

4 Performance Analysis

The improvement of channel gain increases the SNR as well as data rate. Using the Shannon capacity formula, the data rate can be expressed as

$$D_{\text{rate}} = B \log_2\left(1 + \frac{\beta q}{N_O}\right) \tag{8}$$

where B is denoted as the bandwidth of the channel, β is the signal gain, q is the strength of the signal, and N_O is the noise variance.

Improvement of data rate with respect to signal strength is shown in Fig. 3, by varying the channel gain. The enhanced signal with improved data rate requires low power for signal reflection which helps to get better energy efficiency of the reflector surface and the energy efficiency of the reflector surface can be expressed as

$$\text{EE}_{\text{surface}} = \frac{D_{\text{rate}}}{P_{\text{surface}}} \tag{9}$$

where D_{rate} is the data rate of the signal and P_{surface} is the required power for the signal reflector surface.

In communication, the bit error rate (BER) is another important performance parameter that provides the ratio of the number of error bits to the total number transferred bits during communication. Figure 4 shows the BER performance with respect to SNR for direct or only BackCom and RIS-assisted BackCom varying the channel gain and with the increment of SNR, the BER of channel decreases.

5 Conclusion

In this study, we proposed an encrypted RIS-assisted BackCom for secured 6G network and IoT communication which helps to improve the quality as well as performance of signals. The combination of backscatter and RIS technology improves

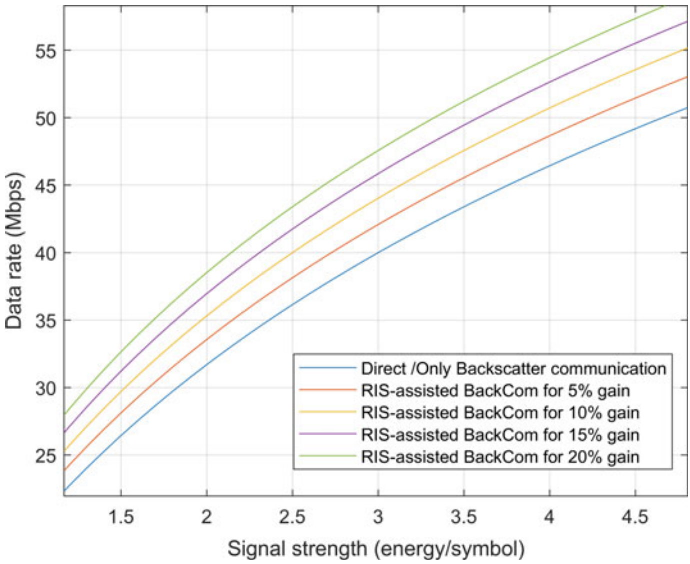


Fig. 3 Data rate versus signal strength varying channel gain

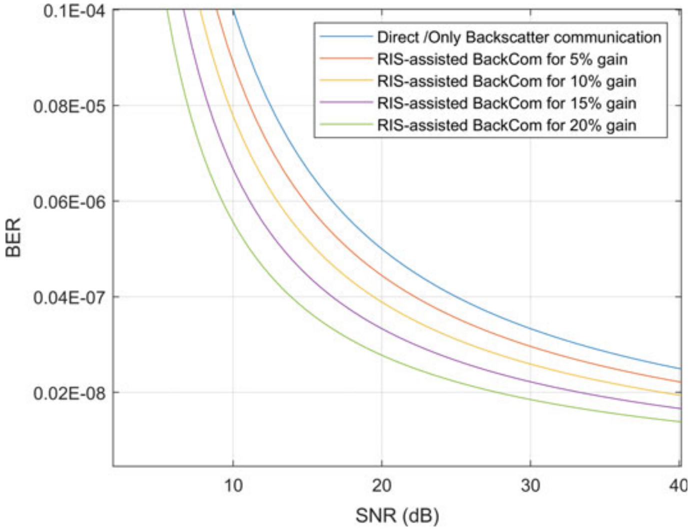


Fig. 4 BER performance analysis with the varying of channel gain

the probability of channel gain than direct link communication. From our analysis, we have shown that the improvement of channel gain increases the strength of the signal as well as data rate, the energy efficiency of reflector surface, and introduced different types of RIS-assisted BackCom topology. In particular, the implementation of RIS-assisted BackCom technology will change a lot in 6G communication using its combined data communication system which will increase the data rate, signal strength, and energy efficiency of the reflecting surface and decrease the BER of transmitted signal which are very fundamental requirements of next generation wireless networks.

References

1. Chowdhury, M.Z., Shahjalal, M., Ahmed, S., Jang, Y.M.: 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions. IEEE Open J. Commun. Soc. **1**, 957–975 (2020)
2. Strinati, E.C., Alexandropoulos, G.C., Sciancalepore, V., Di Renzo, M., Wymeersch, H., Phan-huy, D.-T., Crozzoli, M., D'Errico, R., De Carvalho, E., Popovski, P., Di Lorenzo, P., Bastianelli, L., Belouar, M., Mascolo, J.E., Gradoni, G., Phang, S., Lerosey, G., Denis, B.: Wireless environment as a service enabled by reconfigurable intelligent surfaces: the RISE-6G perspective (2021)
3. Renzo, M.D., Debbah, M., Phan-Huy, D.-T., Zappone, A., Alouini, M.-S., Yuen, C., Sciancalepore, V., Alexandropoulos, G.C., Hoydis, J., Gacanin, H., de Rosny, J., Bounceur, A., Lerosey, G., Fink, M.: Smart radio environments empowered by reconfigurable AI meta-surfaces: an idea whose time has come. J Wireless Com. Network. **2019**, 129 (2019)
4. Di Renzo, M., Zappone, A., Debbah, M., Alouini, M.-S., Yuen, C., de Rosny, J., Tretyakov, S.: Smart radio environments empowered by reconfigurable intelligent surfaces: how it works, state of research, and the road ahead. IEEE J. Sel. Areas Commun. **38**, 2450–2525 (2020)
5. Jameel, F., Ristaniemi, T., Khan, I., Lee, B.M.: Simultaneous harvest-and-transmit ambient backscatter communications under Rayleigh fading. J. Wireless Com. Network **2019**, 166 (2019)
6. Ji, B., Xing, B., Song, K., Li, C., Wen, H., Yang, L.: The efficient BackFi transmission design in ambient backscatter communication systems for IoT. IEEE Access **7**, 31397–31408 (2019)
7. Fara, R., Ratajczak, P., Huy, D.-T.P., Ourir, A., Di Renzo, M., De Rosny, J.: A prototype of reconfigurable intelligent surface with continuous control of the reflection phase (2021)
8. Wu, Q., Zhang, R.: Towards smart and reconfigurable environment: intelligent reflecting surface aided wireless network (2019)
9. Zhao, W., Wang, G., Atapattu, S., Tsiftsis, T.A., Tellambura, C.: Is Backscatter link stronger than direct link in reconfigurable intelligent surface-assisted system? IEEE Commun. Lett. **24**, 1342–1346 (2020)

Activity Classification from First-Person Office Videos with Visual Privacy Protection

Partho Ghosh, Md. Abrar Istiak, Nayeeb Rashid, Ahsan Habib Akash, Ridwan Abrar, Ankan Ghosh Dastider, Asif Shahriyar Sushmit, and Taufiq Hasan

Abstract With the advent of wearable body cameras, human activity classification from First-Person Videos (FPV) has become a topic of increasing importance for various applications, including life-logging, law enforcement, sports, workplace, and health care. One of the challenging aspects of FPV is its exposure to potentially sensitive objects within the user's field of view. In this work, we developed a visual privacy-aware activity classification system focusing on office videos. We utilized a Mask R-CNN with an Inception-ResNet hybrid as a feature extractor for detecting and later blurring out sensitive objects (e.g., digital screens, human face, paper) from the videos. We incorporate an ensemble of Recurrent Neural Networks (RNNs) with ResNet, ResNeXt, and DenseNet-based feature extractors for activity classification. The proposed system was trained and evaluated on the FPV office video dataset which is a subset of the BON [1] vision dataset for office activity recognition (2021) and includes 18 classes made available through the IEEE Video and Image Processing (VIP) Cup 2019 competition. On the original unprotected FPVs, the proposed activity classifier ensemble reached an accuracy of 85.078% with precision, recall, and F1 scores of 0.88, 0.85, and 0.86, respectively. The performances were slightly degraded on privacy-protected videos, with accuracy, precision, recall, and F1 scores at 73.68%, 0.79, 0.75, and 0.74, respectively.

Keywords Activity classification · Visual privacy protection · First person video

---

P. Ghosh · Md. A. Istiak · N. Rashid · A. H. Akash · R. Abrar · A. G. Dastider
Department of Electrical and Electronic Engineering (EEE), Bangladesh University of Engineering and Technology (BUET), Dhaka 1205, Bangladesh

A. S. Sushmit · T. Hasan (✉)
mHealth Research Group, Department of Biomedical Engineering (BME), Bangladesh University of Engineering and Technology (BUET), Dhaka 1205, Bangladesh
e-mail: taufiq@bme.buet.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_12

1 Introduction

As body-worn cameras are becoming more ubiquitous, automatic logging of human activity from FPVs has become a topic of increasing interest. At present, FPV or PoV (point-of-view) cameras are mostly utilized by athletes, motor drivers, and police officers [2, 3]. Generally, it is not feasible to manually analyze large numbers of such videos. Automatically processing FPVs can thus be beneficial in several major applications, including activity logging, law enforcement, search and rescue missions, inspections, home-based rehabilitation, and wildlife observation [4]. As the Augmented Reality (AR) glasses become mainstream, the availability of FPV data is expected to increase while also raising associated concerns for security and privacy of the users [5].

Activity classification can be done in two different modalities: one using sensor data and the other using video data. Recent developments in the deep learning field have made tremendous progress in classifying activity in sensor data. In [6], they have described such a state-of-the-art method, and all of the other methods follow a similar path. One of the major problems with working with sensor data is that the amount of activity becomes very limited. On the other hand, automatic activity classification from videos has been undertaken using two major approaches in the past. Traditional methods generally extract hand-engineered features from the videos to train a machine learning algorithm for classification. Frequently employed features include average pooling (AP), robust motion features (RMF), and pooled appearance features (PAF) [7, 8]. In recent times, deep neural networks are being used to learn the features and subsequently classify the activities. One of the effective methods involves using a Convolutional Neural Network (ConvNet) with Long Short-term

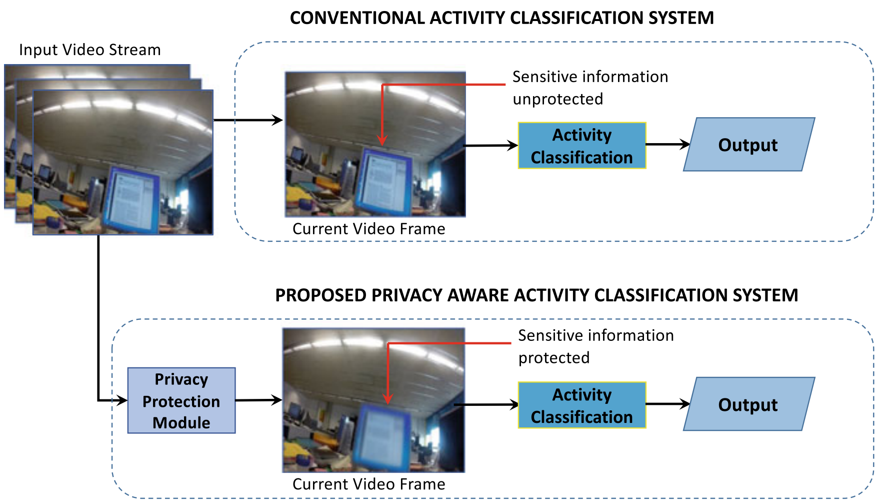


Fig. 1 Conventional versus proposed privacy-aware activity classification systems

Memory (LSTM) [9] units as back-ends [10, 11]. In this approach, the ConvNet learns the front-end image features, whereas the LSTMs recognize the temporal features within the videos that are relevant for human activities. Another well-known method found in literature consists of using a 3D ConvNet to recognize the actions [12]. 3D ConvNets are functionally similar to 2D ConvNets, except that they incorporate Spatio-temporal filters [13] and thus do not specifically require the LSTM layers. However, none of the previous work in the area of activity classification using FPVs has addressed the issue of privacy. For visual privacy protection, recently developed iPrivacy [14] approach protects the most significant privacy-sensitive object via blurring. Serdar et al. proposed a reliable and reversible privacy protection scheme based on false color [15] which affects the whole frame's content. Some of the most popular ways to protect visual privacy are silhouette, scrambling, in-painting [16], and filtering [17]. Latest research on action recognition with visual privacy protection proposed a compression-based coding scheme named VSCS [18] method for privacy protection. It employs a coding scheme to compress images, but VSCS frames lose all content of the frame except the main content which yields visually distracting results.

In this work, we develop a privacy-aware activity classification framework for office videos. In most cases, the personal information contained within FPV videos does not carry useful features for detecting the user's activity. Accordingly, we use a deep learning model to identify the sensitive regions of the video and make them unintelligible visually (e.g., blurred). Next, these privacy-protected videos are used to train deep learning models for activity recognition using the FPV dataset of office activities made available for the IEEE VIP Cup 2019 [7] (Fig. 1). Performance comparisons between activity classification from original vs. privacy-protected videos are performed to demonstrate the effectiveness of the proposed framework.

2 Dataset

The dataset we used in this study is a part of a bigger dataset named BON—Egocentric Vision Dataset for Office Activity Recognition [1], and this subset was provided through the IEEE VIP Cup 2019 competition. The dataset was collected using a chest-mounted GoPro Hero3+ Camera with 1280×760 pixels resolution and a frame rate of 30 fps [7]. Human activities can be broadly categorized as (i) human to human activity, (ii) human to object activity, and (iii) ambulatory activity. The BON dataset contains a total of 18 activity in classes that include all of these broad categories. The dataset and distribution of the 18 classes are summarized in Fig. 2. The figure shows that the class-imbalance problem is a major issue in the dataset that we need to address.

During the competition, the dataset was released in two phases, 400 and 932 videos, respectively. We found label noise in 32 videos through manual annotation, and these are excluded. The remaining videos from the two phases are used to prepare our training and test set containing 873 and 364 videos, respectively.

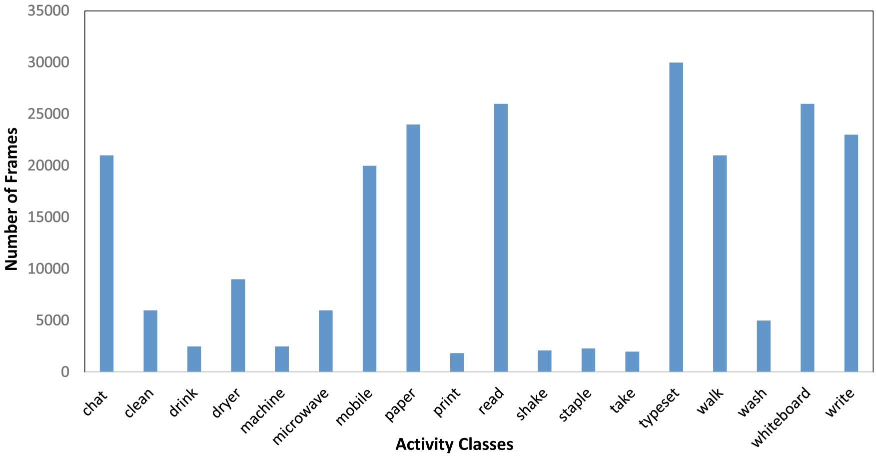


Fig. 2 Bar plot showing the number of video files available for each activity class. The uneven distribution of data in different classes illustrates the issue of class imbalance

Table 1 Training/test splits in the sub-datasets used

Sub-datasets	# videos in training	# videos in test
Original	873 original	364 original
Blurred	873 blurred	364 blurred
Mixed	873 org. + 873 blurred (1746 total)	364 org. + 364 blurred (728 total)

To train the privacy-aware system, we prepared three different sub-datasets for training and testing our deep learning models. We refer them to as the (i) *original*, (ii) *blurred*, and (iii) *mixed* sub-datasets (Table 1). The *original* set contains the original videos from the BON dataset, *blurred* set contains visual privacy-protected videos, and the *mixed* subset contains videos from both subsets (i) and (ii) in an equal amount. The process of generating visual privacy-protected videos in the *blurred* set are described in the following sections.

3 Proposed Architecture

3.1 Pre-processing

Analyzing the histogram of the average brightness of the video frames (mean value of each image pixel), we observed that some videos have a significantly higher brightness compared to others (e.g., data provided in the "Oxford" folder). Thus, we applied gamma correction [19] to the images to normalize this effect. Next, we resize the video frames and normalize them to fit them in the particular deep learning pre-trained networks, Wide ResNet [20], ResNeXt [21], and DenseNet [22], as described later in Sect. 3.3.

3.2 Visual Privacy Protection Module

In this section, we describe the proposed visual privacy protection and security enhancement module that addresses the visual privacy concerns in the BON dataset. The framework consists of two steps: (i) identifying the sensitive regions from a video frame and (ii) making these parts unintelligible by blurring [23, 24].

Identification of Sensitive Regions First, we manually screened the dataset for objects that are prone to privacy violations of the users or may contain visually sensitive information that can lead to a security breach. In the data, we found a total of 7 objects that may contain sensitive information: (i) digital screen, (ii) laptop, (iii) mobile, (iv) book, (v) person, (vi) keyboard, and (vii) toilet/urinal. We used the Mask R-CNN [25] network to identify these objects in our video frames [26].

We utilize the Mask R-CNN approach as it performs instance segmentation [27] to produce a mask for the sensitive objects that may not be of uniform shape (e.g., a person). Traditional methods that generate a bounding-box [28, 29] are not suitable for our application since it will result in the blurring of a larger box-shaped region as compared to the actual sensitive object. This may degrade the activity classifier performance. Examples of object localization (bounding-box) and instance segmentation for an image frame of the activity class "Chat" are shown in Fig. 3a and b, respectively.

Different feature extractor networks can be utilized with a Mask R-CNN model, including Inception-V2 [24], ResNet50 [31], ResNet101 [31], and Inception-ResNet-V2 [24] with Atrous convolution. The performance of these models for object segmentation has been compared on the Common Objects in Context (COCO) dataset [32] with respect to the Mean Average Precision (mAP) metric [33, 34]. These

(a) Object localization (b) Instance segmentation

Fig. 3 Example image frames illustrating object localization and instance segmentation for a video from obtained from the class "Chat" of the BON dataset

Table 2 Performance of different feature extractor-based Mask R-CNN models on the COCO dataset for object segmentation

Model name	mAP
Mask R-CNN ResNet101 Atrous [30]	33
Mask R-CNN Inception-V2	25
Mask R-CNN ResNet50 Atrous	29
Mask R-CNN Inception-ResNet-V2 Atrous [30]	**36**

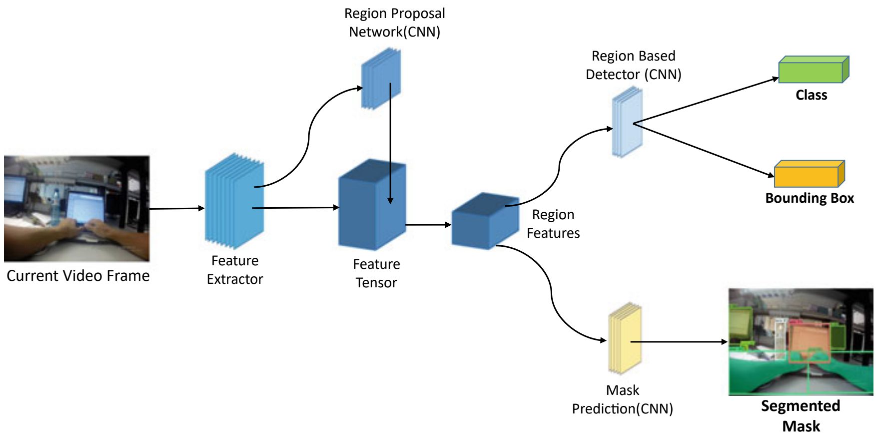


Fig. 4 Model architecture for identification of visually sensitive regions from the video frames. A Mask R-CNN model with an Inception-ResNet-V2 Atrous feature extractor is used

results are available online in the Tensorflow Github repository.[1] According to these mAP scores reproduced in Table 2, Inception-ResNet-V2 with Atrous convolution model provides the best performance on the COCO dataset. Thus, we select this model for our sensitive object segmentation. Since the BON dataset does not contain any object mask labels, it was not possible to objectively evaluate the performance of the object segmentation module on this data. However, we validated the performance of the mask detection model based on visual observation on a subset of BON frames.

The final architecture of the sensitive region segmentation is shown in Fig. 4. We used a pre-trained Inception-ResNet Hybrid-based Mask R-CNN model trained on the COCO dataset. First, we classify between 80 objects available in the COCO dataset. Next, we selected seven (7) potentially sensitive objects, as previously described. Finally, the frames are converted to RGB from BGR and are resized according to the required input dimension of the activity classifier network.

Protecting the Sensitive Objects After identifying the visually sensitive regions from the video frames, a Gaussian filter is used to blur the regions where the sensitive

[1] https://github.com/tensorflow.

(a) Identified mask (b) Sensitive regions blurred

Fig. 5 Example of generated masks using the Mask R-CNN network using the Inception-ResNet-V2 with Atrous convolution model as a feature extractor. The video frame was labeled as the "Typeset" class

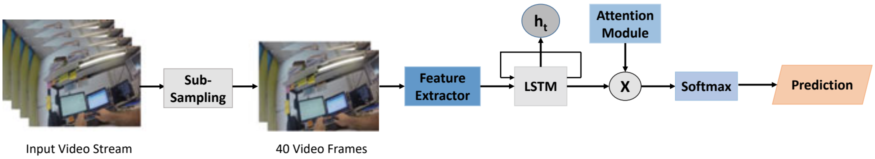


Fig. 6 Proposed activity classification framework excluding the privacy protection module

objects are detected. An example video frame from the "chat" class is shown in Fig. 5a along with the detected mask using the Mask R-CNN network, while Fig. 5b shows the same frame with visual sensitive objects blurred (hands, screens, and keyboard).

3.3 Activity Classification Model

The proposed activity classification module consists of a 3-channel single stream network inspired by [35]. The network is illustrated in Fig. 6. The authors of [35] used AlexNet [36] as the feature extractor with two uni-directional LSTM layers for temporal sequence modeling. In contrast, we propose to utilize an ensemble of Wide ResNet [20], ResNeXt [21], and DenseNet [22] feature extractors and a single-layer bi-directional LSTM with framewise attention. The proposed framework is detailed in the following subsections.

Video Frame Selection We sub-sampled video frames to reduce the computational load. Regardless of video duration, we extracted 40 frames. To select these 40 frames, we first determine a sampling period by dividing the total number of frames in the video by 40 and taking the floor of the result. Then, we randomly took the first frame in the range of 1 to the sampling period. After selecting the first frame, we jump the sampling period and take the next frame, and so on until all 40 frames are taken. In this way, we were able to fit almost all the video information into these 40 frames. We experimented with different values except 40, but 40 frames per video produced the best result. These image frames are then gamma-corrected and normalized prior to loading into the model.

Feature Extraction Layer We experimented with different feature extractor models including DenseNet121 [22], DenseNet161 [22], ResNet50 [31], ResNet101 [31], ResNet152 [31], Inception-V3 [37], Inception-ResNet-V2 [24], ResNeXt [21], and Wide ResNet101 [20]. In these experiments, we used pre-trained models on ImageNet [38]. We froze the feature extractor and trained all the remaining layers of our model.

LSTM Layer We pass the video frames to a bi-directional single-layer LSTM network [39]. Our LSTM layer had an input dimension of 512, returned by the feature extractor, and a hidden state size of 1024.

Framewise Attention We hypothesize that, for activity classification, all the sampled video frames are not of equal temporal importance. Some frames are temporally more important to activity recognition than others. Accordingly, we use a framewise attention module with a softmax activation function after the LSTM layer to give more weightage to the feature vectors corresponding to the frames that are more important to recognize the activity happening in the video.

Fully Connected Layers The output of the framewise attention module is passed to the fully connected layer, followed by batch normalization and a softmax decision layer.

Training Scheme The proposed activity classification network is trained for 20 epochs with the Adam optimizer [40] using cross-entropy as the loss function. To address the class-imbalance issue in the data, we implemented a balanced training scheme where every mini-batch contained an equal number of samples from each class. We oversampled the underrepresented classes while creating these balanced mini-batches. We begin training with an initial learning rate of 0.001 and implemented a learning rate scheduler with a patience value of 5 and with a decay factor of 1/10 to facilitate model convergence.

4 Experimental Evaluation

4.1 Performance of Activity Classification Models

In our activity classification network, we implemented a variety of feature extractors and compared their results in different settings. During our experiments, we observed that each model tested performs best at a specific frame size. The top-performing models and their performance on the *original*-test data is shown in Table 3.

4.2 Performance of Model Ensemble

We propose to use an ensemble of the four classifiers described in the previous section. Ensemble weights are calculated based on the classwise F1 score of each model. We first compare the performance of the ensemble using models trained and tested on different data subsets. The results provided in Table 4 show that the ensemble model trained only on the *original*-training set performs poorly on the *blurred*-test set. However, training on *mixed* dataset improves the *blurred*-test performance with a drop of accuracy in the *original*-test condition. We address this trade-off by adding fine-tuned models in our final ensemble.

4.3 Model Tuning and Final Ensemble

To address the issue of performance degradation on the *blurred* test, we select our models trained on the *original* sub-dataset and use transfer learning to fine-tune them using the *blurred* (privacy-protected) set. We prepare the final ensemble using the four original models mentioned in Table 3 and the proposed fine-tuned versions. Prediction time of ensemble classifier is extensively discussed in the IEEE VIP Cup final evaluation [41]. On 5-s samples of each task, this ensemble classifier took 94.42 s on average.

Table 3 Performance comparison of different networks on the *original*-test dataset

Model	Resolution	Accuracy (%)
1. ResNeXt	248 × 248	75.92
2. DenseNet	512 × 512	74.87
3. Wide ResNet101 + Attention	324 × 324	79.84
4. Wide ResNet101	324 × 324	75.39

Table 4 Accuracy of the proposed model ensemble on different test sets while using different training sets

Ensemble training data	Accuracy (%)	
	Original-test	Blurred-test
Original-train dataset	85.07	68.8
Mixed-train dataset	82.72	75.8

Table 5 Overall performance of the final ensemble classifier

Sub-dataset	Precision	Recall	F1 Score	Accuracy (%)
Original	0.88(±0.1)	0.85(±0.1)	0.86(±0.1)	85.08(±11.4)
Blurred	0.79(±0.2)	0.75(±0.2)	0.74(±0.2)	73.68(±19.97)

4.4 *Results and Discussion*

The results of the final ensemble and classwise F1 scores are shown in Table 5 and Fig. 7, respectively. From the results, we first observe that the class imbalance did not affect our system performance as we used balanced-batch training technique. As an illustrative example, from Fig. 2 we observe that the class "typeset" included a significantly larger amount of data compared to the "take" class. The final classwise F1 scores depicted in Fig. 7 show that the system performed better in detecting the minority class. Secondly, we observe that due to the visual privacy protection the performance of the initial network dropped quite significantly. The performance gap between original videos and protected videos were about 20%. To address it, we train our network using the visual protected videos. That improved the result on protected videos but degraded the results on original videos and again created a huge performance gap in between them. We achieve best results when we trained the model on original videos and fine-tuned those models using the protected videos. In this way, the performance gap between the original videos and visual privacy-protected videos was significantly reduced and we reached a more generalized approach. On average, the performance metrics in Table 5 degraded about 10% from their original values due to visual privacy protection, demonstrating the effectiveness of the proposed privacy-aware activity classification system.

5 Conclusions

In this work, we have developed a privacy-aware activity classification from FPV videos using an ensemble of deep learning models. The visual privacy protection module utilized pre-trained networks to identify sensitive regions within the video frames and performed Gaussian blurring on those pixels. The system has been trained using balanced mini-batches to effectively address the issue of class imbalance in the training data. The ensemble of models trained on unprotected videos and later fine-tuned on privacy-protected videos provided the best overall performance in both conditions.

Acknowledgements We want to thank the department of BME and Brain Station 23 (Dhaka, Bangladesh) for supporting this research. The NVIDIA Corporation donated the TITAN Xp GPU used for this work.

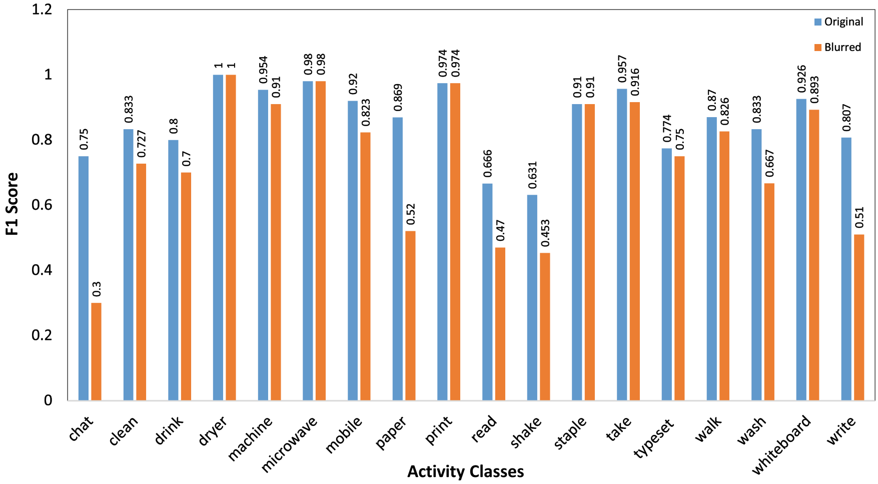


Fig. 7 F1 score distribution on *original*-test and *blurred*-test dataset videos across the class labels

References

1. Tadesse, G.A., Bent, O., Weldemariam, K., Istiak, M.A., Hasan, T.: Bon-egocentric vision dataset for office activity recognition. (2021) [Online]. Available: https://dx.doi.org/10.21227/ehzr-w794
2. Ryoo, M.S., Matthies, L.: First-person activity recognition: what are they doing to me?. In: Proceeding of IEEE CVPR (2013), pp. 2730–2737
3. Chen, H., Li, H., Song, A., Haberland, M., Akar, O., Dhillon, A., Zhou, T., Bertozzi, A.L., Brantingham, P.J.: Semi-supervised first-person activity recognition in body-worn video (2019). ArXiv preprint arXiv:1904.09062
4. Tadesse, G.A., Cavallaro, A.: Visual features for ego-centric activity recognition: a survey. In: Proceeding 4th ACM WearSys. ACM, pp. 48–53 (2018)
5. Roesner, F., Kohno, T., Molnar, D.: Security and privacy for augmented reality systems. Comm. ACM **57**(4), 88–96 (2014)
6. Sikder, N., Ahad, M.A.R., Nahid, A.-A.: Human action recognition based on a sequential deep learning model. In: 2021 Joint 10th International Conference on Informatics, Electronics and Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision and Pattern Recognition (icIVPR). IEEE, pp. 1–7 (2021)
7. Abebe, G., Catala, A., Cavallaro, A.: A first-person vision dataset of office activities. In: Proceeding IAPR MPRSS-HCI. Springer, pp. 27–37 (2018)
8. Sun, S., Kuang, Z., Sheng, L., Ouyang, W., Zhang, W.: Optical flow guided feature: a fast and robust motion representation for video action recognition. In: Proceeding IEEE CVPR, pp. 1390–1399 (2018)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceeding IEEE CVPR, pp. 2625–2634 (2015)
11. Kahani, R., Talebpour, A., Mahmoudi-Aznaveh, A.: A correlation based feature representation for first-person activity recognition. Multimedia Tools Appl. **78**(15), 21 673–21 694 (2019)

12. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal **35**(1), 221–231 (2012)
13. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceeding of IEEE CVPR, pp. 6299–6308 (2017)
14. Yu, J., Zhang, B., Kuang, Z., Lin, D., Fan, J.: iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. IEEE Trans. Inf. Forens. Secur. **12**(5), 1005–1016 (2016)
15. Çiftçi, S., Akyüz, A.O., Ebrahimi, T.: A reliable and reversible image privacy protection based on false colors. IEEE Trans. Multimedia **20**(1), 68–81 (2017)
16. Paruchuri, J., Cheung, S.-C., Hail, M.: Video data hiding for managing privacy information in surveillance systems. EURASIP J. Inf. Secur. **2009**, 1–18 (2009)
17. Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., Vincent, L.: Large-scale privacy protection in google street view. In: IEEE 12th International Conference on Computer Vision. IEEE **2009**, 2373–2380 (2009)
18. Liu, J., Zhang, R., Han, G., Sun, N., Kwong, S.: Video action recognition with visual privacy protection based on compressed sensing. J. Syst. Arch. **113**, 101882 (2021)
19. Marcu, G., Zhong, J., Swen, S.: Dynamic gamma correction. Dec. 21 2006, uS Patent App. 11/153,959
20. Zagoruyko, S., Komodakis, N.: Wide residual networks (2016). ArXiv preprint arXiv:1605.07146
21. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceeding IEEE CVPR, pp. 1492–1500 (2017)
22. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceeding IEEE CVPR, pp. 4700–4708 (2017)
23. Xu, W., Cheung, S.-C.S., Soares, N.: Affect-preserving privacy protection of video. In: Proceeding ICIP. IEEE, pp. 158–162 (2015)
24. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Prof. 31st AAAI (2017)
25. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceeding IEEE ICCV, pp. 2961–2969 (2017)
26. Lim, J.J., Salakhutdinov, R.R., Torralba, A.: Transfer learning by borrowing examples for multiclass object detection. In: Proceeding NIPS, pp. 118–126 (2011)
27. Romera-Paredes, B., Torr, P.H.S.: Recurrent instance segmentation. In: Proceeding ECCV. Springer, pp. 312–329 (2016)
28. Long, Y., Gong, Y., Xiao, Z., Liu, Q.: Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Trans. Geosci. Remote **55**(5), 2486–2498 (2017)
29. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans. Pattern Anal. **39**(1), 189–203 (2016)
30. Rosenfeld, A., Zemel, R., Tsotsos, J.K.: The elephant in the room (2018). ArXiv preprint arXiv:1808.03305
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceeding IEEE CVPR, pp. 770–778 (2016)
32. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: Proceeding ECCV. Springer, pp. 740–755 (2014)
33. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale (2018). ArXiv preprint arXiv:1811.00982
34. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceeding IEEE CVPR, pp. 7310–7311 (2017)
35. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W.: Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access **6**, 1155–1166 (2017)

36. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceeding NIPS, pp. 1097–1105 (2012)
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceeding IEEE CVPR, pp. 2818–2826 (2016)
38. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceeding IEEE CVPR, IEEE, pp. 248–255 (2009)
39. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. ArXiv preprint arXiv:1508.01991 (2015)
40. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). ArXiv preprint arXiv:1412.6980
41. Tadesse, G.A., Bent, O., Marcenaro, L., Weldemariam, K., Cavallaro, A.: Privacy-aware human activity recognition from a wearable camera: highlights from the IEEE video and image processing cup 2019 student competition [sp competitions]. IEEE Sig. Proc. Magaz. **37**(3), 168–172 (2020)

A Blockchain-Based Secured Land Record System Using Hyperledger Fabric

Md. Anwar Hussen Wadud, Tahmid Ahmed, Alhaj Hossen, Md. Mahedi Hasan, Md. Aminul Islam, and Md. Hasibur Rahman

Abstract In Bangladesh, scarcity of land and fast population growth is putting a strain on the land-man ratio. The property ownership registration system in Bangladesh is incomplete and insufficient. As a result, various government agencies handle various documents, and bureaucracy flaws enable organized crime. Considering these, we are proposing a method where we have come up with a Blockchain-based solution. We build a Blockchain-based system to ensure that individuals are not deceived under any circumstances. In our system, where land data will be safe and secure, data synchronization and transparency will be available, as well as ease of access, irreversible record management, and rapid and low-cost quick solutions. We designed a modern architecture to digitally store land records using Blockchain-based Hyperledger Fabric to ensure land security. This proposed model has bâŁ˜een presented by private Blockchain, taking into account the technological expertise and authority of the people and government. Finally, we compare our proposed architecture with the existing land records management model. Our system provides more security and data privacy than other models and saves both money and time in our daily lives. This will ensure that our system is transparent and acceptable in all directions.

Keywords Land registration · Blockchain · Hyperledger fabric · Privacy · Security

1 Introduction

Land registration is a system that keeps track of who owns something and what rights they have over it. Blockchain is a remarkable new technology that can provide any agency's systems trust, integrity, and accessibility. A database, or Blockchain,

Md. A. H. Wadud
Department of CSE, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh

Md. A. H. Wadud (✉) · T. Ahmed · A. Hossen · Md. M. Hasan · Md. A. Islam · Md. H. Rahman
Department of CSE, Bangladesh University of Business and Technology, Dhaka, Bangladesh
e-mail: mwadud@bubt.edu.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_13

is a sort of digital ledger. In most cases, Blockchain accumulates data in groups, sometimes known as blocks, that contain sets of data. Blocks have specific storage capabilities, and when filled, they are chained onto the previous block, establishing a data chain known as the "Blockchain". All additional information added after that newly added block is compiled into a new block, which is then added to the chain after it is filled [1].

Hyperledger Fabric is a private Blockchain. Businesses can use distributed ledger technology without exposing their data to the public by using a private Blockchain. This website is password-protected. The Blockchain allows only known nodes to participate in the procedure [2].

The primary problem in the current system is that data is dispersed across multiple government institutions that are not well-synchronized, enabling hacked persons to modify legal papers. A centralized system will be insufficient to cope with the various land titling frauds in this case [3]. In this research, we evaluate Bangladesh's present land registration process and propose a new framework based on Blockchain technology to improve transparency and dependability [4].

In the past, there was a manual system for land registration. At that time, we faced many problems. This existing system takes a long time. On the basis of Bangladesh, there are some illegal means of registration. However, before this study, no work has been introduced on the digitization of land registration using Hyperledger Fabric framework. In this paper, we are going to introduce the digital version of the land registration system where we use a Hyperledger Framework with Blockchain technology [5].

The present age is the digital and modern age, and technological development is increasing day by day. Everything we need in our daily lives is going digital. As we have analyzed, people have already been recording land with pen in hand which is not going to keep pace with the digital age, and this manual system leaves people facing a lot of harassment. However, to analyze human demand, we want to develop a land record system that is easily accessible to the user. This system is very easy to use for users, and it is very secure for user information. The study work's overall performance is as follows:

- We highlight a current challenge in analyzing human demand in the secured land service.
- We propose a secured land registration model using Hyperledger Fabric-based Blockchain network.
- We reduce manual involvement by implementing actual property title assignments and ensuring the security of public land data.

The following is how the rest of this article is organized: Sect. 2 discusses the land verification process in Bangladesh, and Sect. 3 discusses the related works of the land record management system. Section 4 briefly explains Blockchain technology, smart contract, and Hyperledger Fabric. Also, Sect. 4 discusses system architecture and implementation of our recommended framework. Finally, this study concludes with future guidance in Sect. 5.

2 Land Verification Process in Bangladesh

2.1 *Existing Methods for Registering Property in Bangladesh*

In Bangladesh, anybody has the legal right to process an immovable property of land. The general public in Bangladesh is ignorant of the complex record of rights (ROR) transfer processes, which require hundreds of papers handled by several government departments and take a long time to complete. As a result, intermediaries use every administrative and legal gap they can find and use bribes to alter papers in order to harass ordinary people. This, in turn, generates years of legal wrangling and is the primary source of civil disputes in Bangladesh [3]. In Fig. 1, we discussed the current working procedure of the land record management system in Bangladesh.

First of all, meet both the buyer and the seller for oral discussions. Then, the land documents are examined and verified. The buyer and seller “Agreement to Sell” and get it notarized. If the needed buyer wants to loan for land. Buyer and seller go to a lawyer then verify their agreement. By the lawyer, the document with land money has to be recorded within the stamp paper. Buyer transfers money through the sub-registration office to the seller. The sub-registration office provides a money receipt

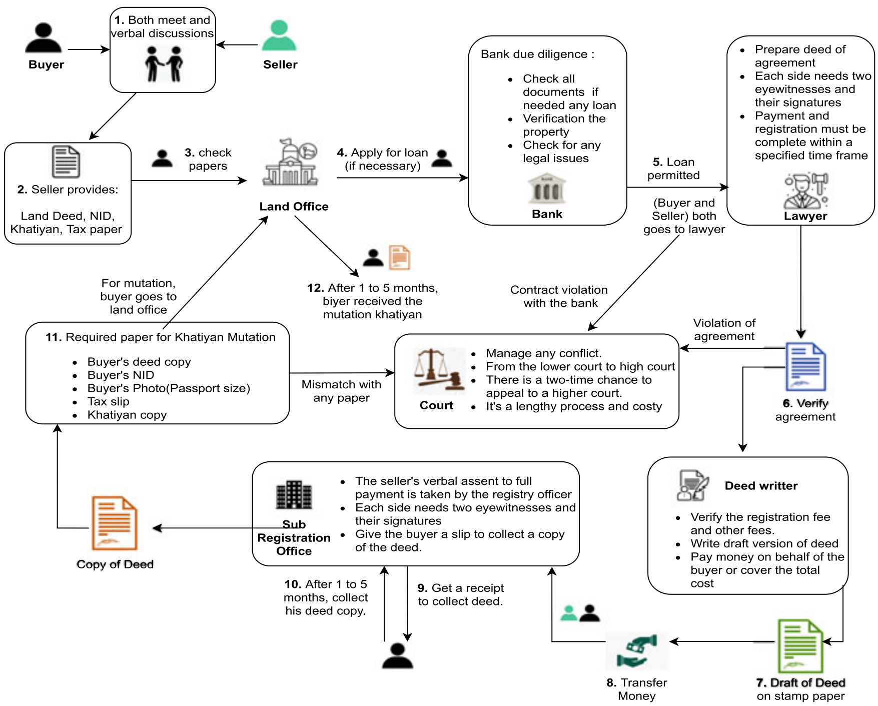


Fig. 1 Existing methods for registering property in Bangladesh

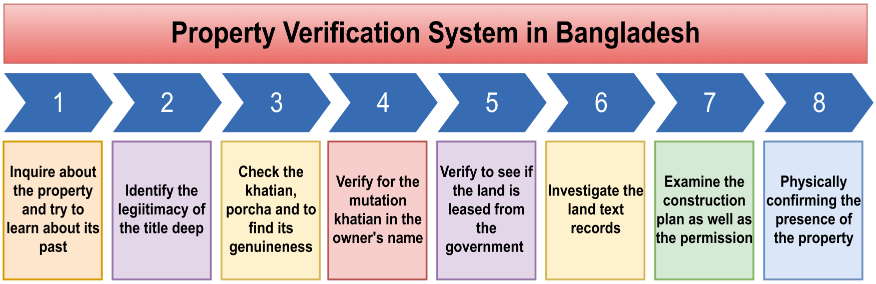


Fig. 2 Property verification system in Bangladesh

to the buyer. After one to five months, he collected copies of the documents from the sub-registration office. After obtaining all information, the sub-register office sends it to the land office for mutation purposes. And after one to five months, buyer receives the mutation Khatiyan through the land office. This is how the existing system works.

2.2 *How to Verify if a Property Is Presently Owned in Bangladesh?*

Property ownership disputes are quite prevalent in Bangladesh. Property documents are readily faked and untrustworthy. If a person is not careful while purchasing a property, he or she may encounter issues, including a possible dispute with the property's ownership at a later date. Verifying land ownership in Bangladesh, on the other hand, is a time-consuming task. A buyer can check the ownership of a property by following the procedures outlined in Fig. 2.

2.3 *Complications with the Present Land Registration System*

There are several problems in our current system.

- **Verification of Ownership**: Land registration agencies across the world confront a number of problems, one of which is ensuring owner validation.
- **Historical Record of Ownership**: These assets do not have a recorded owning history in many cases. When negotiating with unknown persons, getting access to an asset's entire historical record of owning (for example, a piece of land) enhances confidence.
- **Illegally Obtained Land Sales**: Property may be sold without the owners' or insurers' permission, resulting in economic loss.
- **Getting Late for Ownership of Land Registration**: Land registration and post-pones in the transfer of owning on paper are time-consuming, lasting over the

month. Erroneous property value might result in erroneous tax or insurance premiums.
- **Fails to Recognize a Fraud**: Present paper-based or technological records are ineffective in preventing identification fraud and identity theft, which can lead to unlawful transactions.

3 Related Works

Md. Sakibul Islam et al. [4] proposed digital land registration model, where they used Ethereum, a public and permissioned Blockchain to build their model. In their article, the privacy question is answered to a large extent by integrating the land registry on Blockchain. This will be the most suited and also give infinite processing capabilities, but there has been no earlier work on private Blockchains such as Hyperledger Fabric. Due to the complexity of the technology, its maturity level, and unusual first implementation that does not emphasize the true benefit of Blockchain, consumers have diverse opinions and attitudes about the technology [6]. They explained the earliest Blockchain technologies were public implementations for cryptocurrencies. Kazi et al. [3] suggested a hybrid Blockchain-based system that assures stakeholder coordination, data transparency, ease of access, and controls immutable transaction records in this research. Given the government's and general public's technological immaturity, they suggested a three-stage Blockchain adoption strategy that begins with public Blockchain and progresses to a large-scale full hybrid Blockchain. To establish an unchangeable and genuine database for land purchasers and owners, certificates were specially formatted and encrypted on the public Bitcoin Blockchain [7]. The system proposed in [8] this paper follows a decentralized approach of private Blockchain with various consensus algorithms for validating each transaction of land between the new owner and old owner without the involvement of third parties. For a safe and trustworthy land register system, a Blockchain-based architecture is being developed by using Ethereum [9]. Table 1 shows the recent works of the different proposed models with our proposed model. Mukne et al. [1] propose a single-level anti-corruption model where anyone can access the system with admin permission. This model enables Internet protocol Edge Services (IPES). Consider column two—Alam et al. [3] propose a three-phase digitized model where anyone can access it easily. This model enables Internet protocol Edge Services (IPES). Consider column three—Thakur et al. [10] propose a single phase ownership transparent model where anyone cannot access the system because of the private data. This model disables the Internet protocol Edge Services (IPES). Consider our proposed model as an archive digitization model where anyone can access the system with admin permission. This model enables Internet protocol Edge Services (IPES).

Table 1 Compression of different proposed models

Keyword	Mukne et al. [1]	Alam et al. [3]	Thakur el al. [10]	Our proposed model
Propose year	(2019)	(2020)	(2019)	(2021)
Country	India	Bangladesh	India	Bangladesh
Challenge	Anti-corruption	Digitized	Ownership transparent	Archive digitization
Proposed model	One phase	Three phase	One phase	Incremental two phase
Blockchain architecture	Permissioned	Public and hybrid	Privet or public	Permission
IPES	Enabled	Enabled	N/A	Enabled
Experimental result	Moves phase one to two	Prototype	N/A	Compared with benchmark

4 System Architecture

4.1 Proposed Model Workflow

Distributed network implementation in real life is the main challenge for us. Because land registration system we can use two types of data. One of them is public and the other one is private data [11]. Public data is available for every user, but private data can see only specific users such as Ac land and surveyor. So, our main goal is to define who sees the public data and who sees the private data. Fig. 3 shows the system architecture of the proposed model. When a user comes to take our service, we will first ask him/her to register. If he/she has registered earlier, he/she will need to log in and enter our system. After the login, an option will come in front of him/her by going to the option called Check Details and he/she will have to be given details input of the land which he/she is willing to purchase. As soon as he/she inputs, our system will show him/her all the details of the land where the area, quantity, landowner, and everything will be there. This will prevent the sale of fake land. If all the information is correct, an option will be shown that says "Wants to registration". Clicking on this option will require filling up a form that will provide all rules and conditions and land information, land price, etc. After submitting this form, a confirmation letter will go to the landowner. If the owner accepts the form, then show the button which is "pay the registration fee", otherwise, there show a massage which is the owner not accepting your request. When paying the registration fee, then it transfers to the admin panel where all admin checks the papers. If all papers are well decorated as per the rules and regulations, then it is asking for the payment for land price. The transaction will be credited to Bangladesh Bank. This money will be handed over to the land owner after a certain period of time. Through this, we will also be able to ensure the security of payments. After transaction, a confirmation message would

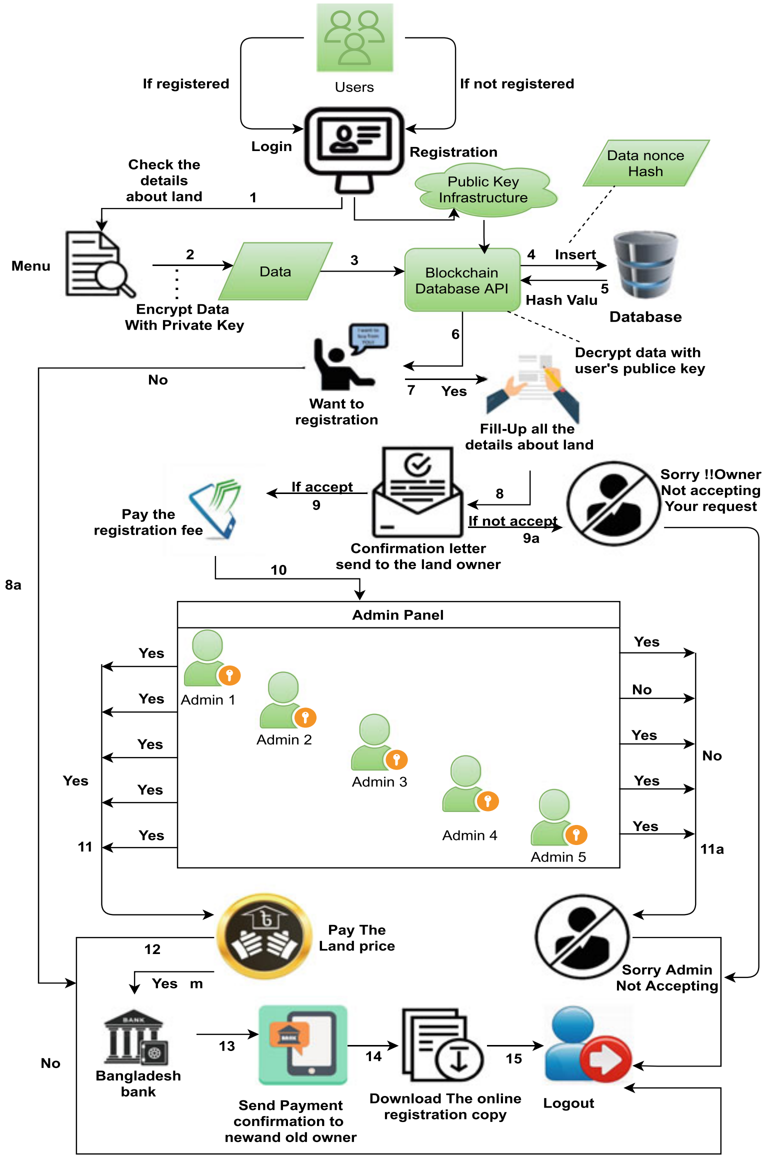


Fig. 3 Proposed system framework for land registration system

send both the new owner and the old owner. Finally, the new owner will be allowed to download an online copy of the registration along with the government seal. This is how the system works.

4.2 Working Strategy of Blockchain Technology

Blockchain is a decentralized digitized ledger of transactions that is copied and distributed throughout Blockchain's whole computer network structures [12–23]. Every block of the chain comprises a number of transactions, and when a new transaction takes place on the Blockchain, a record of that transaction is recorded to each user's ledger [9]. Blockchain is a distributed ledger system in which data is stored using a hash, which is a cryptographic signature that cannot be changed [24]. If hackers wanted to harm a Blockchain system, they'd have to change each block in the chain across all distributed versions.
(1) Smart Contract: A smart contract is a type of code (written in Go, node.js, or Java) that enforces contract processing between parties using Hyperledger Fabric. Smart contracts use data from the Blockchain to automate corporate processes and workflows [25]. An idea of smart contract automation is monitoring an SLA and subsequently committing a reduction to the ledger for the given IT solution if it is violated.
(2) Hyperledger Fabric: Hyperledger Fabric is a private Blockchain. A private Blockchain enables businesses to use distributed ledger technology without providing their data to the public. This is a permissioned Blockchain, which means that only recognized nodes can join the network [2]. The Fabric has several components. The major components of fabric are listed below.

1. **Membership Service Provider(MSP)**: Membership Service Provider (MSP) is a component that attempts to encapsulate the membership operation architecture.
2. **Client**: Clients are apps that suggest transactions over the network on behalf of a user. To connect with a network, the client employs a Fabric SDK.
3. **Peer**: The state and the ledger transact a node to maintain a copy. The ordering service sends blocks to a peer, who maintains the state and ledger.
4. **Ordering Service**: The ordered service offers orderers a connection that ensures delivery for the Hyperledger Fabric.

(3) Hyperledger Fabric Work Flow: Hyperledger Fabric is a permissioned Blockchain network created by organizations that want to form a consortium. Each Blockchain network member organization is responsible for ensuring that its peers are ready to join the network. Fig. 4 shows the step by step workflow of the Hyperledger Fabric network.

- A transaction request is initiated by a participant in the member organization using the client application.
- The transaction invocation request is broadcast by the client application to the Endorser peer.
- To confirm the transaction, the Endorser peer examines the certificate information and the others. The chaincode (i.e., smart contract) is then executed, and the endorsement replies are returned to the client. As part of the endorsement response, the Endorser peer provides transaction approval or rejection.

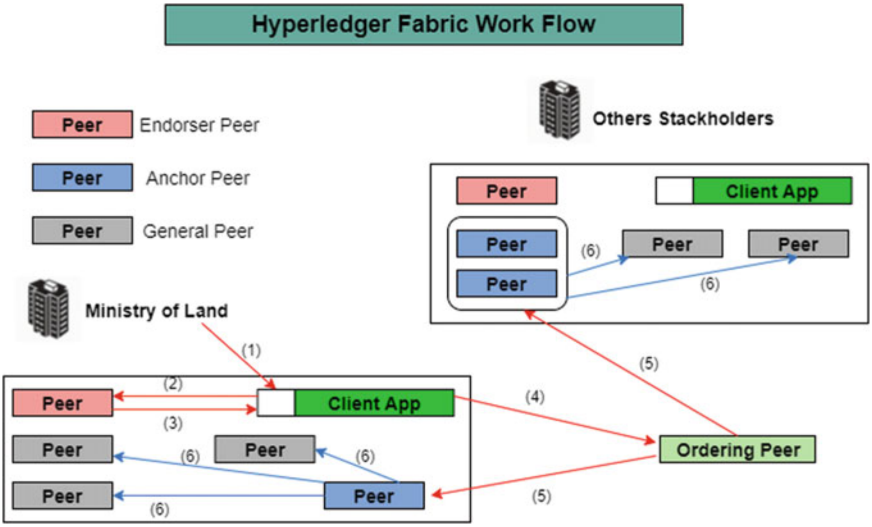


Fig. 4 Hyperledger Fabric workflow

- The client now transmits the authorized transaction to the Orderer peer, who will organize it correctly and put it in a block.
- The transaction is placed in a block by the Orderer node, then sends the block to the Anchor nodes of the Hyperledger Fabric network's various member organizations.
- Anchor nodes then disseminate the block to the rest of their organization's peers. The newest block is then updated in each peer's local ledger.

5 The Implementation of Blockchain in Land Record Registration

Represents the class diagram of the private Blockchain platform's smart contract development objects (user profile, add a property, transfer property, user owned land, and transaction) as shown in Fig. 5. Personal information, addresses are stored in the user's NID. With this NID user information, there is an interface for retrieving currently owned land and whether the user exists. A unique ID will be provided to someone while adding the land details to the database. ID ensures ownership of land. The landowner can promote his own land with the unique ID mentioned. When anyone can buy and sell any land or property then can use transfer property and transaction section otherwise, he will not use this system. User owner land section confirms the existence of the landowner and provides information. Similarly, there are necessary identifier information, multiple ownership support, ownership transfer process. On the other hand, a land registration ensures the location of the object and the identifier information. Each object is bound to the user of the Blockchain.

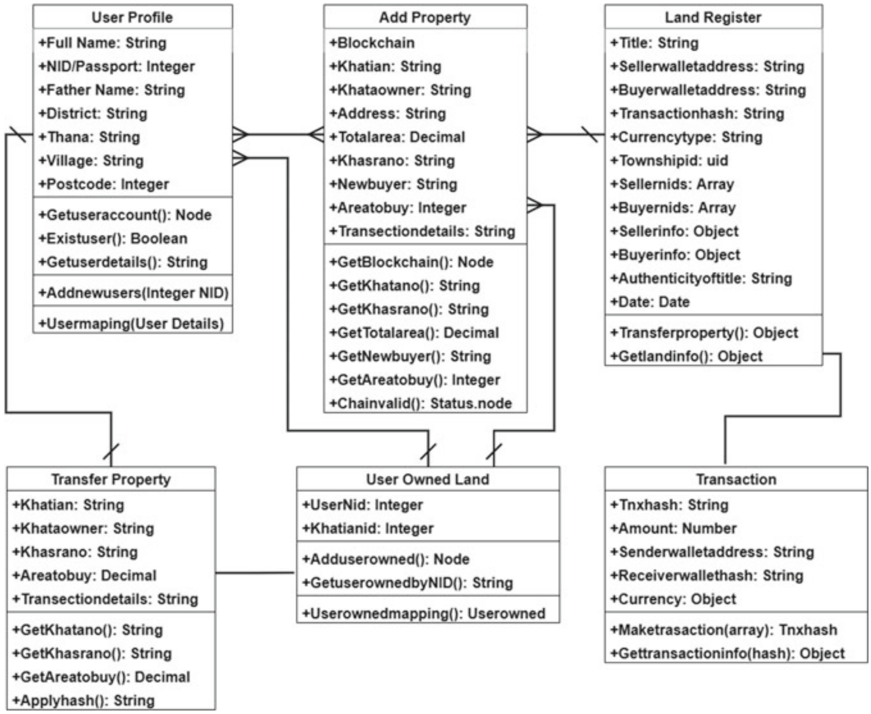


Fig. 5 Class diagram of development land registration system

6 Experimental Analysis

In this part, we discuss the proposed system's implementation. There are two types of Blockchain technology, as previously stated: public Blockchain and private Blockchain. We have chosen the private Blockchain for implementation of the proposed system because our proposed system only requires a few nodes so it is a cost-effective approach. In this paper, a variety of Blockchain implementations have been deployed for constructing Blockchain-based applications [12–23]. We utilized the Hyperledger Fabric framework (Linux Foundation, Hyperledger Fabric 3) for Ubuntu (Linux version) to develop the suggested system since it is a private Blockchain implementation with good reliability and a stable community. We also utilized a Hyperledger Composer that appears to support the Hyperledger Fabric framework and run-time (Linux Foundation, Hyperledger Composer 4). It is a suite of tools for creating Blockchain-based solutions quickly. Other tools used to implement the proposed system involve: (a) Client URL (a library for sending data using different protocols), (b) Docker and Docker-Compose (used to construct and operate numerous Blockchain network containers), (c) PHP, (Hypertext Preprocessor), and (d) NodeJs (used to create smart contracts).

Table 2 System environment

Hardware	Configuration
Operating system	Ubuntu Linux 18.04.1 LTS
CPU	Single vCPU @ 2.00GHz
Memory	32 GB
Hyperledger Fabric	Version 1.2
Docker-Compose	Version 1.5.2
Oracle VirtualBox	Version 6.1.22
Docker	Version 1.2.1

Table 3 Comparison of existing works with proposed work

Key terms	Mukne et al. [1]	Alam et al. [3]	Thakur et al. [10]	Proposed model
User validation	✓	✓	✓	✓
Identity management	✓	**X**	✓	✓
Data monitoring	✓	✓	✓	✓
Decentralized access	✓	✓	✓	✓
Availability	✓	**X**	**X**	✓
Flexibility	**X**	**X**	**X**	✓

We perform the offered model utilizing Hyperledger Composer and Fabric. Through our analyses, we assume that the user data is recovered of the JSON and demanded data by using the Rest User, for instance, Postman server. Each server was created on the Virtual EC2 instance On AWS, which serves as Ubuntu Linux 18.04.1, 32 GB RAM, and single VCPU @ 2.00 GHz on the similarly local PC as the description of the summary in Table 2. To develop the entire architecture, we applied the Hyperledger Composer playground. We utilized Hyperledger Fabric (v 1.2) Linux foundation with hosted an open-source project.

We have used Blockchain technology to ensure user validation used in our project where each customer ensures validation by matching the land ledger number is shown Table 3. Identity management, data monitoring and availability, as well as decentralized access, flexibility and speedy, and low-cost quick solutions, will all be accessible in our system, which will keep land data safe and secure. Our proposed system has a decentralized network with an immutable transaction history for a forge-proof system that can also keep track of all papers online. Lastly, our best reason for all models is the time factor. Other models need maximum time, whereas our model takes less time for each task. Additional transactions can be done through our model in the short term. As we see, the graph of Fig. 6 shows that each set needs less time to complete the task. But as the number of tasks increases, the number of executions will also increase. Although our model always needs less time than others. When the number of tasks is 0–20, then our model completes the transaction

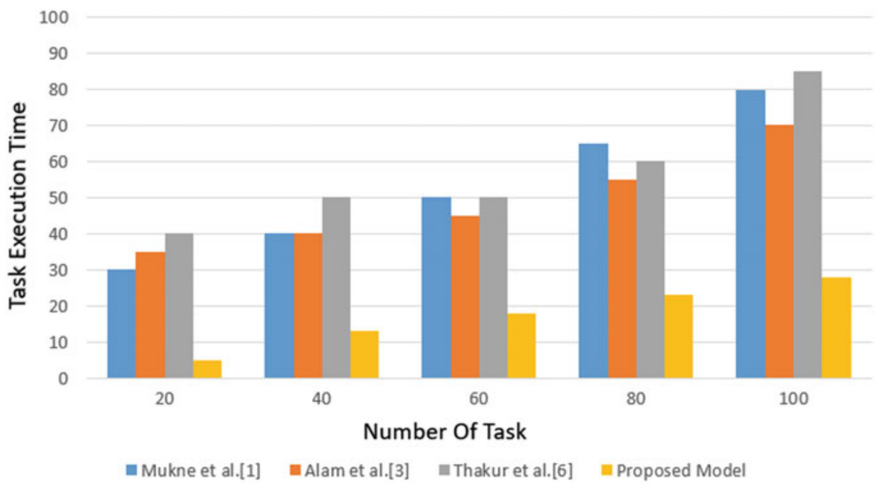


Fig. 6 Tasks number versus execution time

in 5 s, whereas other models require 30–45 seconds. Similarly, when the number of tasks is 20–40, 40–60, 60–80, and 80–100, our model takes 13, 18, 23, and 28 s, respectively, whereas other models take up to 40–85 seconds.

7 Conclusion

In Bangladesh, registered land maintenance and its frequent updating have proven to be a difficult undertaking. The general public has lost faith in the current systems. Many are not sure if anyone else's name is on the land in their name. And many people cannot be sure who the real owner of the land is when they buy land? We have to go to the land office to register and face various problems. That is why we have introduced a Blockchain-based online land record system. In our system, people can register land very quickly without any problem. Updating our system with accurate information can quickly register land. One of the many benefits is that no one can change the transaction history. There will be no doubt about it, and no one will be able to show fake records even if they want to. Thus, this Blockchain-based system proves that all matters of transfer of land ownership is very protective, and it is essential for a corrupt country like us.

References

1. Mukne, H., Pai, P.., Raut, S., Ambawade, D.: Land record management using hyperledger fabric and IPFS. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–8. IEEE (2019)
2. Islam, Md.S. : Land verification system using blockchain technology in Bangladesh (2019)
3. Alam, K.M., Rahman, J.M.A., Tasnim, A., Akther, A.: A blockchain-based land title management system for Bangladesh. J. King Saud Univ.-Comput. Inf, Sci (2020)
4. Islam, Md.S., Iqbal, F.S., Islam, M.: A novel framework for implementation of land registration and ownership management via blockchain in Bangladesh. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 859–862. IEEE (2020)
5. Zhang, D., Zhang, Z., Chen, L., Li, S., Huang, Q., Liu, Y.: Blockchain technology hyperledger framework in the internet of energy. In: IOP Conference Series: Earth and Environmental Science, vol. 168, pp. 012043. IOP Publishing (2018)
6. Papantoniou, C., Hilton, B.: Enterprise solutions criteria in the age of geoblockchain: land ownership and supply chain. In: Proceedings of the 54th Hawaii International Conference on System Sciences, pp. 5307 (2021)
7. Yaga, D.., Mell, P.., Roby, N., Scarfone, K.: Blockchain technology overview. *arXiv preprint* arXiv:1906.11078 (2019)
8. Sharma, R.., Galphat, Y., Kithani, E., Tanwani, J., Mangnani, B., Achhra, N.: Digital land registry system using blockchain. Available at SSRN 3866088 (2021)
9. Graglia, J.M., Mellon, C.: Blockchain and property in 2018: at the end of the beginning. Innovations: Technol. Gov. Globalization **12**(1-2), 90–116 (2018)
10. Thakur, V., Doja, M.N., Dwivedi, Y.K., Ahmad, T., Khadanga, G.: Land records on blockchain for implementation of land titling in India. Int. J. Inf. Manage. **52**, 101940 (2020)
11. Nasir, Q., Qasse, I.A., Abu Talib, M., Nassif, A.B.: Performance analysis of hyperledger fabric platforms. Sec. Commun, Netw (2018)
12. Wadud, Md.A.H., Amir-Ul-Haque Bhuiyan, T.M., Uddin, Md.A., Rahman, Md.M.: A patient centric agent assisted private blockchain on hyperledger fabric for managing remote patient monitoring. In: 2020 11th International Conference on Electrical and Computer Engineering (ICECE), pp. 194–197 (2020)
13. Biswas, M., Al Faysal, J., Ahmed, K.A.: Landchain: a blockchain based secured land registration system. In: 2021 International Conference on Science and Contemporary Technologies (ICSCT), pp. 1–6. IEEE (2021)
14. Mukherjee, P.P., Boshra, A.A., Ashraf, M.M., Biswas, M.: A hyper-ledger fabric framework as a service for improved quality e-voting system. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 394–397. IEEE (2020)
15. Al-Amin, S., Sharkar, S.R., Kaiser, M.S., Biswas, M.: Towards a blockchain-based supply chain management for e-agro business system. In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering, pp. 329–339. Springer (2021)
16. Hossain, Md.J., Wadud, Md.A.H., Rahman, A., Ferdous, J., Alam, Md.S., Haque Bhuiyan, T.M.A.U., Firoz Mridha, M.: A secured patient's online data monitoring through blockchain: an intelligent way to store lifetime medical records. In: 2021 International Conference on Science Contemporary Technologies (ICSCT), pp. 1–6 (2021)
17. Biswas, M.., Kaiser, M.S., et al.: Drlas: digital record keeping in land administration system relying on blockchain. In: Proceedings of Sixth International Congress on Information and Communication Technology, pp. 965–973. Springer (2022)
18. Hossain, Md.J., Wadud, Md.A.H., Hdm-chain, Md.A.: A secure blockchain-based healthcare data management framework to ensure privacy and security in the health unit. In: 2021 5th International Conference on Electrical Engineering and Information and Communication Technology (ICEEICT) (2021)

19. Rabbi, Md., Hradoy, P.M., Islam, Md.M., Islam, Md.H., Akter, M.Y., Biswas, M.: Bls: bank loan sanction using blockchain authenticity, transparency and reliability. In: 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), pp. 1–5. IEEE (2021)
20. Noman, Md.A.A., Hossain, Md.J., Kalimulla, Md.M., Azad, S., Wadud, Md.A.H.: An intelligent application for preventing the counterfeit medicines through a distributed blockchain. In: 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI) (2021)
21. Ahmed, S., Shome, A., Biswas, M.: Dbst: a scalable peer-to-peer distributed information system supporting multi-attribute range query. In: 2021 International Conference on Science and Contemporary Technologies (ICSCT), pp. 1–6. IEEE (2021)
22. Biswas, M., Akhund, T.Md.N.U., Ferdous, M.J., Kar, S., Anis, A., Shanto, S.A.: Biot: blockchain based smart agriculture with internet of thing. In: 2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4), pp. 75–80. IEEE (2021)
23. Ahamed, S., Siddika, M., Islam, S., Anika, S., Anjum, A., Biswas, M.: Bps: blockchain based decentralized secure and versatile light payment system. Asian J. Res. Comput. Sci. 12–20 (2021)
24. Ali, S., Wang, G., White, B., Cottrell, R.L.: A blockchain-based decentralized data storage and access framework for pinger. In: 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 1303–1308. IEEE (2018)
25. Watanabe, H., Fujimura, S., Nakadaira, A., Miyazaki, Y., Akutsu, A., Kishigami, J.: Blockchain contract: securing a blockchain applied to smart contracts. In: 2016 IEEE International Conference on Consumer Electronics (ICCE), pp. 467–468. IEEE (2016)

HealthBlock: A Secured Healthcare System Using Blockchain

Md. Mahfujur Rahman, Md. Nur Amin Sifat, Mostafizur Rahman, Mushfiqur Rahman, Shamim Al Mamun, and M. Shamim Kaiser

Abstract Blockchain technology enables a distributed and decentralized environment with no more central authority. To increase the accuracy of electronic healthcare records (EHRs) and establish a secured patient-centric approach, Blockchain can be a smart solution in this case. Blockchain is a distributed ledger technology that allows for the secured transfer of medical records while a transaction is created in the network. As Blockchain is decentralized and transparent if one user tampers with medical records of transactions, all other nodes/participants would cross-reference each other and easily pinpoint the node/participants with the wrong information. In this paper, we proposed a smart contact-based Blockchain technology and implemented it in the Hyperledger framework to make the healthcare system more confidential and secure. Moreover, Smart contracts give us more security while transacting data and reduce the entire transaction cost, and also make faster transaction speed.

Keywords Blockchain · Hyperledger fabric · Smart contract

1 Introduction

In the healthcare system, large information is produced throughout the world with multidimensional medical services information. There is an increasing interest in digitalizing healthcare systems by governments and related industry sectors, partly evidenced by various initiatives taking place in different countries and sectors. However, healthcare data can be faltered to share delicate clinical information. Therefore, patient data and related huge scope information mining have become a troublesome test. Blockchain innovation is considered to have the potential for securing medical data. Blockchain is a decentralized framework, in which importing the information isn't put away in a focal area and it isn't being constrained by a single

Md. M. Rahman · Md. N. A. Sifat · M. Rahman · M. Rahman
Daffodil International University, Dhaka, 1207, Bangladesh

S. A. Mamun · M. S. Kaiser (✉)
Jahangirnagar University, Savar, Dhaka, 1342, Bangladesh
e-mail: mskaiser@juniv.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_14

person. Blockchain is decentralized and no central position controls the substance added to the Blockchain [1]. Rather, the passages given to the Blockchain are settled upon in a distributed system utilizing different agreement conventions. The attribute of Blockchain is steadiness. It is difficult to erase sections in the wake of being acknowledged onto the Blockchain because of the conveyed record, put away over various hubs. Moreover, the chance of namelessness is an engaging trademark used in numerous Blockchains [1].

Blockchain has the potential to solve problems like clinical data confidentiality, privacy, sharing, and storage in the healthcare sector. Interoperability is a key necessity in EHRs because it allows healthcare practitioners and patients to interchange data or information precisely, efficiently, and consistently so that it can be shared throughout the environment and disseminated by different hospital systems. Costs are key factors in healthcare infrastructure and software, putting enormous strain on global economies. Blockchain technology is improving patient outcomes, optimizing business processes, patient data management, enhancing compliance, lowering costs, and enabling better use of healthcare-related data in the healthcare sector [2].

One critical step in making the healthcare system smarter and improving the quality of healthcare services and users' experiences is the capacity to share healthcare data without jeopardizing users' privacy and data security. The goal of this work is to present a secure healthcare data sharing system that addresses privacy concerns.

Our contribution in this paper is as follows:

- A brief description of the existing Blockchain model is delineated in the healthcare system.
- A smart contract-based Blockchain architecture with Hyperledger fabric is proposed in the healthcare eco-system.
- A simple application is implemented using proposed model.

The rest of the paper is organized as follows. In Sect. 2, Related works are stated of Blockchain. Challenges for Blockchain in the healthcare system is stated in Sect. 3. In Sect. 4, we proposed a smart contract-based framework implementation with hyperledger fabric of Blockchain in the healthcare system. In Sect. 4, we discussed the results and implemented proposed model. Finally, the conclusion discussed in Sect. 6.

2 Related Works

There are many applications [3–7] using Blockchain technology that has been proposed to verify certificates, secure patient data, secured healthcare system, smart secured land administration system, supply chain management, and more. In this part, some traditional approaches based on Blockchain technology for secured healthcare systems are presented. Azaria et al. [8] proposed framework gives secure access to the data of the patients. The clients know about the clinical history and patients are being taken in certainty changes should be possible if necessary. Benefits are given to users to get rights to the framework on what kind of understatement will appear to the Blockchain excavators. An association is made between the foundation with the current information stockpiling. The proposed framework accomplishes decentralization of clinical records in a made sure way. The structure utilized the brilliant agreement instrument and POW-based accord calculation to approve another square in their Blockchain-based framework. Cirstea et al. [9] describe the Blockchain innovation followed by clinical applications (MedBlock). MedBlock changes the general clinical framework proficiency. Anybody can offer involvement in medicines, disease, sensitivities, and so forth that also help to framework gather information and doctors, researchers and drugstores effectively get to these records. The authors utilize the sha-2 and sha-3 encryption techniques for security purposes. Yaqoob et al. [10] describe that in medical services, patients need security, realness, straightforwardness of exchange appropriately. Blockchain technology handles the issue of computerized trust by safely recording significant data in an open space. The authors give a few impediments of this model: these are a danger to legitimacy, a danger to ID and determinations of essential examinations, a danger to data extraction, and so forth. McGhin et al. [11] depict insight concerning shrewd contact, extortion location, and personality confirmation. The authors acquainted us with some significant issues that they guarantee that current Blockchain applications don't top off some medical services novel prerequisites, however, scientists attempt to beat these issues. Tanwar et al. [12] proposed an entrance control strategy algorithm for improving information accessibility between human services suppliers. This is a Hyperledger-based electronic healthcare services record sharing framework that utilizes the idea of a chain code. Some valuable ideas which can measure Blockchain execution, for example, inactivity, throughput, Right transfer transaction (RTT) they attempt to utilize here precisely. In addition, contrasted with conventional Electronic social insurance record frameworks, which use customer server design, they proposed a framework that utilizes Blockchain for improving the effectiveness and security of human services frameworks. The authors propose two algorithms here, one algorithm is an administrator and has full access to a framework. Administrators can add applicable individuals to the framework and have full access to read, write and refresh and expel any client from the framework. Another algorithm is a patient block can log in framework utilizing a private key. The patient can access, read, write his data. If any client is substantial, at that point client can be seen by the patient, clinician, and research facility staff records over a Blockchain organization. Hasselgrena et al. [13] wanted to improve the presence of Blockchain highlights for the social insurance framework, health science, and health instruction. Creators audit a few papers and the greater part of the papers examine the Electronic social insurance framework, individual medicinal services framework utilizing Blockchain innovation. Ethereum and Hyperledger textures are well-known Blockchain innovations. The authors presented us with a record database.

3 Challenges

Uncertainty: The Blockchain idea isn't far reaching yet and there are just a couple of effective activities dependent on this cutting edge innovation right now. That is a significant obstacle since we don't have numerous effective Blockchain models to follow which makes an unsure circumstance.

Storage Capacity: Blockchain inside the healthcare services industry will be made out of clinical records, pictures, archives, and lab reports which require a lot of extra room. Theoretically, every part remembered for the chain would have a total duplicate of the full clinical record of each person in the U.S. also, this volume might surpass the capacity limit of current Blockchain innovation.

Information Proprietorship: Who will claim the medical record information? Who will allow consent to share it? That sort of organized office or procedure hasn't been built up yet.

Cost: The expense of building up and keeping up a medicinal services Blockchain is obscure yet and nobody can truly consider this innovation without thinking about its costs early.

Rules and Guidelines: There are no standards accessible to address the utilization of Blockchain in the human services industry. It is additionally uncertain concerning how new arrangements with respect to human services Blockchain will adjust to current security guidelines like the HIPAA demonstration.

4 Proposed Methodology

In this section, we proposed a model through smart contract approach and how medical data is secured and fair transaction through implementing the model in future. Proposed model construct smart representations of existing medical records that are kept on the network within individual nodes using smart contracts. We create contracts that provide metadata about record ownership, permissions, and data quality. In this system, related stakeholder such as patient, doctor, nurse, lab technician, and pharmacy as shown in Fig. 1 registered as a node and create transaction among them through smart contract which is implemented by hyperledger framework. The whole transaction process and its data are saved in the blocks of Blockchain through smart contracts.

To preserve performance and economic sustainability, all medical record data is saved in local database storage such as couchdb, and the hash of the data and committed to the chain. The Blockchain smart contract system has been used to create and implement several medical functionalities in the healthcare system. This services can be a medical prescriptions to the treatment of severe diseases and related procedures, such as surgery patients' treatment procedures. The goal of creating these medical smart contracts is to make it easier for doctors, patients, and healthcare organizations

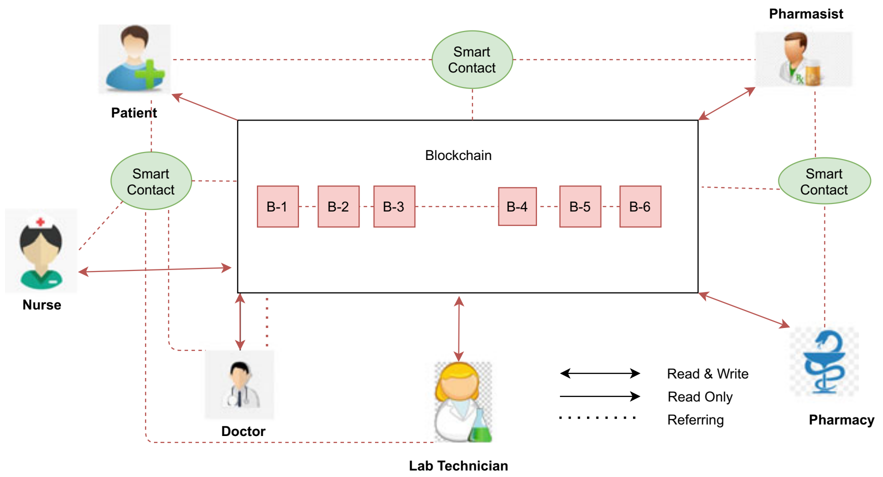


Fig. 1 Smart contract-based proposed model for healthcare system

to overcome bottlenecks in the administrative process. A doctor can writes a prescription for a patient and uses a smart contract to add it to the patient's medical records. The pharmacy then has access to this prescription via a Blockchain smart contract that has been approved by both the primary doctor and the patient. In this way, the smart contract in our suggested approach can manage all transactions. Participants in our suggested model include a doctor, a patient, a nurse, a lab technician, and a pharmacist, as depicted in Fig. 1. Participants can use Blockchain smart contracts to communicate information with labs, doctors, emergency clinics, and other partners.

5 Implementation of Proposed Model

The proposed application, HealthBlock is tested, executed and evaluated on the hyperledger fabric network. All transactions executed using the deployed smart contract in the hyperledger fabric network. Certificate authority provides the notion of identity for the participants who will generate transaction on the Blockchain. So, this identity will be digital certificate and participants will do signed that transaction using this digital certificate and submit them to the Blockchain so that they authenticate with the Blockchain that they are a legitimate user and it also ensures that they get the right access privileges on the Blockchain for the transaction they are performing. Endorser peer checks the transaction by verifying the requester's certificate and roles, as well as executing the Smart Contract and simulating the transaction's conclusion. For the Hyperledger Fabric network, the orderer peer serves as the primary communication route. The Orderer peer/node is in charge of ensuring that the Ledger state is maintained consistently across the network. The block is created by the orderer peer, who then distributes it to all peers Fig. 2.

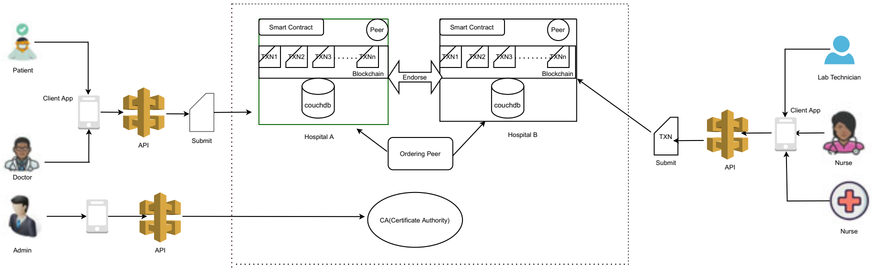


Fig. 2 Implementation procedure in hyperledger fabric of healthcare system


```
id "patient_patient@gmail.com"

{
 "id": "patient_patient@gmail.com",
 "key": "patient_patient@gmail.com",
 "value": {
  "rev": "4-7bd655baaec7cf635a3b5f51dad41019"
 },
 "doc": {
  "_id": "patient_patient@gmail.com",
  "_rev": "4-7bd655baaec7cf635a3b5f51dad41019",
  "docType": "patient",
  "email": "patient@gmail.com",
  "key": "patient_patient@gmail.com",
  "name": "Mr. Patient",
  "pAddress": "Jamalpur",
  "password": "f814f514bdba559aa943fd3c5ada38a97d91bb02a399a8f6a2db691b00dacb22",
  "patientDiagnosis": "Fever",
  "~version": "CgMBTAA="
 }
}
```


Fig. 3 Create patient transaction


```
{
    "key": "patient_patient@gmail.com",
    "email": "patient@gmail.com",
    "password": "f814f514bdba559aa943fd3c5ada38a97d91bb02a399a8f6a2db691b00dacb22",
    "name": "Mr. Patient",
    "pAddress": "Jamalpur",
    "patientDiagnosis": "Fever",
    "docType": "patient"
}
```


Fig. 4 Create patient UI response

Figure 3 represents the transaction of patient creation which will be saved on the Blockchain network. Every transaction has unique doc type, id or key so that later it can trace every successful transaction data.

Figure 4 represents the patient creation client side API response. In client side user can send this data and save it in the blockchain.

```
id "doctor_doctor@gmail.com"

{
 "id": "doctor_doctor@gmail.com",
 "key": "doctor_doctor@gmail.com",
 "value": {
  "rev": "5-b7d9c42536e9b17b1697ec30ca0d2b19"
 },
 "doc": {
  "_id": "doctor_doctor@gmail.com",
  "_rev": "5-b7d9c42536e9b17b1697ec30ca0d2b19",
  "dname": "Mr. Doctor",
  "docType": "doctor",
  "email": "doctor@gmail.com",
  "key": "doctor_doctor@gmail.com",
  "password": "f814f514bdba559aa943fd3c5ada38a97d91bb02a399a8f6a2db691b00dacb22",
  "qualification": "MBBS",
  "salary": "10000",
  "~version": "CgMBTQA="
 }
}
```


Fig. 5 Create doctor transaction


```
{
    "key": "doctor_doctor@gmail.com",
    "email": "doctor@gmail.com",
    "password": "f814f514bdba559aa943fd3c5ada38a97d91bb02a399a8f6a2db691b00dacb22",
    "dname": "Mr. Doctor",
    "qualification": "MBBS",
    "salary": "10000",
    "docType": "doctor"
}
```


Fig. 6 Create doctor UI response

Figure 5 represents the doctor creation transaction. Every transaction have unique doc type, id or key which can trace every successful transaction data.

Figure 6 represents doctor creation client side view. Any doctor can fill his/her information here and register as doctor in Blockchain.

Figure 7 represents the pharmacy creation transaction. Pharmacy can be only one or multiple here. Same as previous transaction every transaction has unique doc type, id or key which can trace every successful transaction data.

Figure 8 represents the pharmacy creation client side view. Any user can easily fill all pharmacy requirement and registered as a pharmacy of this healtcare system. Figures 9, 10, 11, 12, 13, 14 represents the Blockchain transaction via smart contract. All transaction which created in the Blockchain networks will save in couchdb Database which shown in Fig. 15.

```
id "pharmacy_Hospital Pharmacy"

{
 "id": "pharmacy_Hospital Pharmacy",
 "key": "pharmacy_Hospital Pharmacy",
 "value": {
  "rev": "3-b7d86fb6b83a25c7f49d0c5faa43bf39"
 },
 "doc": {
  "_id": "pharmacy_Hospital Pharmacy",
  "_rev": "3-b7d86fb6b83a25c7f49d0c5faa43bf39",
  "docType": "pharmacy",
  "key": "pharmacy_Hospital Pharmacy",
  "pname": "Hospital Pharmacy",
  "~version": "CgMBTwA="
 }
}
```


Fig. 7 Create pharmacy transaction


```
{
    "key": "pharmacy_Hospital Pharmacy",
    "pname": "Hospital Pharmacy",
    "docType": "pharmacy"
}
```


Fig. 8 Create pharmacy UI response

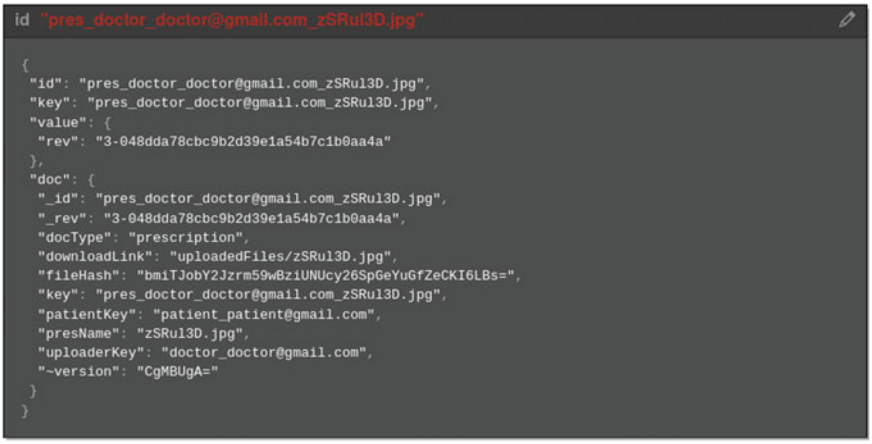

```
id "lab_2c70e12b7a0646f92279f427c7b38e7334d8e5389cff167a1dc30e73f826b683"

{
 "id": "lab_2c70e12b7a0646f92279f427c7b38e7334d8e5389cff167a1dc30e73f826b683",
 "key": "lab_2c70e12b7a0646f92279f427c7b38e7334d8e5389cff167a1dc30e73f826b683",
 "value": {
  "rev": "3-e794920f7e268573c1b915e731a50c88"
 },
 "doc": {
  "_id": "lab_2c70e12b7a0646f92279f427c7b38e7334d8e5389cff167a1dc30e73f826b683",
  "_rev": "3-e794920f7e268573c1b915e731a50c88",
  "docType": "lab",
  "key": "lab_2c70e12b7a0646f92279f427c7b38e7334d8e5389cff167a1dc30e73f826b683",
  "~version": "CgMBUAA="
 }
}
```


Fig. 9 Create lab technician transaction

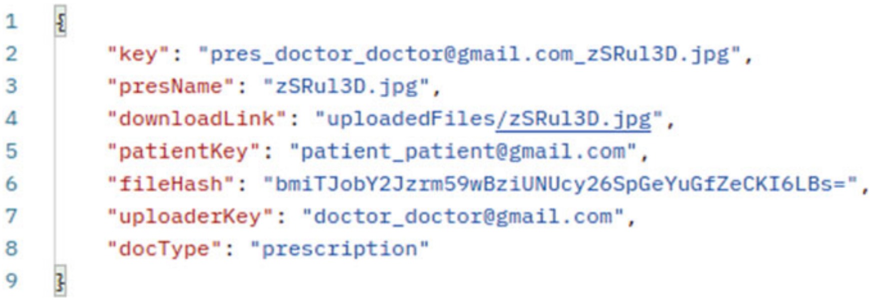

```
{
    "key": "lab_2c70e12b7a0646f92279f427c7b38e7334d8e5389cff167a1dc30e73f826b683",
    "docType": "lab"
}
```


Fig. 10 Create lab technician response



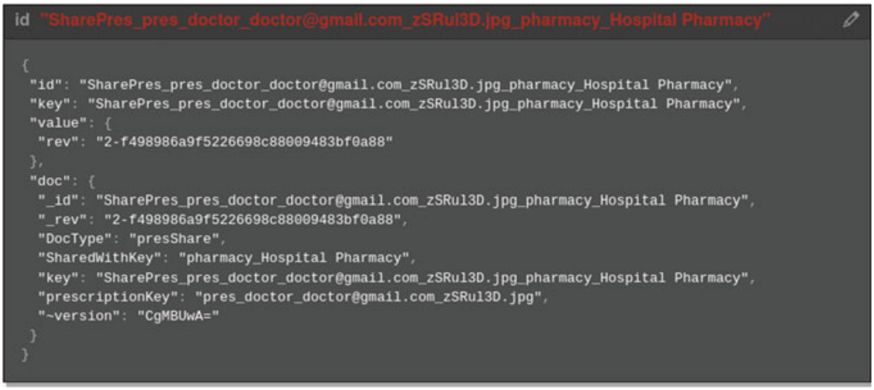
```
id "pres_doctor_doctor@gmail.com_zSRul3D.jpg"
{
 "id": "pres_doctor_doctor@gmail.com_zSRul3D.jpg",
 "key": "pres_doctor_doctor@gmail.com_zSRul3D.jpg",
 "value": {
  "rev": "3-048dda78cbc9b2d39e1a54b7c1b0aa4a"
 },
 "doc": {
  "_id": "pres_doctor_doctor@gmail.com_zSRul3D.jpg",
  "_rev": "3-048dda78cbc9b2d39e1a54b7c1b0aa4a",
  "docType": "prescription",
  "downloadLink": "uploadedFiles/zSRul3D.jpg",
  "fileHash": "bmiTJobY2Jzrm59wBziUNUcy26SpGeYuGfZeCKI6LBs=",
  "key": "pres_doctor_doctor@gmail.com_zSRul3D.jpg",
  "patientKey": "patient_patient@gmail.com",
  "presName": "zSRul3D.jpg",
  "uploaderKey": "doctor_doctor@gmail.com",
  "~version": "CgMBUgA="
 }
}
```


Fig. 11 Create prescription transaction



```
{
    "key": "pres_doctor_doctor@gmail.com_zSRul3D.jpg",
    "presName": "zSRul3D.jpg",
    "downloadLink": "uploadedFiles/zSRul3D.jpg",
    "patientKey": "patient_patient@gmail.com",
    "fileHash": "bmiTJobY2Jzrm59wBziUNUcy26SpGeYuGfZeCKI6LBs=",
    "uploaderKey": "doctor_doctor@gmail.com",
    "docType": "prescription"
}
```


Fig. 12 Create prescription response



```
id "SharePres_pres_doctor_doctor@gmail.com_zSRul3D.jpg_pharmacy_Hospital Pharmacy"
{
 "id": "SharePres_pres_doctor_doctor@gmail.com_zSRul3D.jpg_pharmacy_Hospital Pharmacy",
 "key": "SharePres_pres_doctor_doctor@gmail.com_zSRul3D.jpg_pharmacy_Hospital Pharmacy",
 "value": {
  "rev": "2-f498986a9f5226698c88009483bf0a88"
 },
 "doc": {
  "_id": "SharePres_pres_doctor_doctor@gmail.com_zSRul3D.jpg_pharmacy_Hospital Pharmacy",
  "_rev": "2-f498986a9f5226698c88009483bf0a88",
  "DocType": "presShare",
  "SharedWithKey": "pharmacy_Hospital Pharmacy",
  "key": "SharePres_pres_doctor_doctor@gmail.com_zSRul3D.jpg_pharmacy_Hospital Pharmacy",
  "prescriptionKey": "pres_doctor_doctor@gmail.com_zSRul3D.jpg",
  "~version": "CgMBUwA="
 }
}
```


Fig. 13 Share prescription transaction

```
{
    "key": "SharePres_pres_doctor_doctor@gmail.com_zSRul3D.jpg_pharmacy_Hospital Pharmacy",
    "prescriptionKey": "pres_doctor_doctor@gmail.com_zSRul3D.jpg",
    "SharedWithKey": "pharmacy_Hospital Pharmacy",
    "DocType": "presShare"
}
```


Fig. 14 Share prescription response

_id	key	~version	docType	email
SharePres_pres_d...	SharePres_pres_d...	CgMBWQA=		
doctor_doctor@g...	doctor_doctor@g...	CgMBVQA=	doctor	doctor@gmail.com
lab_2c70e12b7a0...	lab_2c70e12b7a0...	CgMBVwA=	lab	
patient_patient@g...	patient_patient@g...	CgMBVAA=	patient	patient@gmail.com
pharmacy_Hospita...	pharmacy_Hospita...	CgMBVgA=	pharmacy	
pres_doctor_docto...	pres_doctor_docto...	CgMBWAA=	prescription	

Fig. 15 All transaction in couchdb database

6 Conclusion

Exploration on the utilization of Blockchain in medical services is currently settled as a scholastic field, and the number and nature of distributions are expanding quickly. This study tried to find out the current Blockchain research trends in the healthcare system and also find out the vulnerabilities of existing research works on healthcare using Blockchain technology. This study also proposed a smart contract-based secured medical record data storage in decentralized mode. This study is also trying to give a research direction for the researcher for further improvement in the electronic health record system. In the future, we will try to implement this at the production level.

References

1. Darlington, N.: Blockchain for beginners: what is blockchain technology? A step-by-step guide (2016). Accessed 21 Oct 2021
2. How using blockchain in healthcare is reviving the industry's capabilities. Accessed 21 Oct 2021
3. Shawon, S.K., Ahammad, H., Shetu, S.Z., Rahman, M., Hossain, S.A.: DIUcerts DApp: a blockchain-based solution for verification of educational certificates. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE (2021)
4. Meghla, T.I., Rahman, M.M., Biswas, A.A., Hossain, J.T., Khatun, T.: Supply chain management with demand forecasting of covid-19 vaccine using blockchain and machine learning. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE (2021)

5. Biswas, M., Akhund, T.M.N.U., Shamim Kaiser, M.: DRLAS: digital record keeping in land administration system relying on blockchain. In: Proceedings of Sixth International Congress on Information and Communication Technology, pp. 965–973. Springer Singapore, Singapore (2022)
6. Mamun, A.A., Jahangir, M.U.F., Azam, S., Shamim Kaiser, M., Karim, A.: A combined framework of InterPlanetary file system and blockchain to securely manage electronic medical records. In: Advances in Intelligent Systems and Computing, pp. 501–511. Springer Singapore, Singapore (2021)
7. Shamim Kaiser, M., Mamun, S.A., Mahmud, M., Tania, M.H.: Healthcare robots to combat COVID-19. In: COVID-19: Prediction, Decision-Making, and its Impacts, pp. 83–97. Springer Singapore, Singapore (2021)
8. Azaria, A., Ekblaw, A., Vieira, T., Lippman, A.: MedRec: using blockchain for medical data access and permission management. In: 2016 2nd International Conference on Open and Big Data (OBD). IEEE (2016)
9. Cirstea, A., Enescu, F.M., Bizon, N., Stirbu, C., Ionescu, V.M.: Blockchain technology applied in health the study of blockchain application in the health system (II). In: 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE (2018)
10. Yaqoob, S., Murad, M., Talib, R., Dawood, A., Saleem, S., Arif, F., Nadeem, A.: Use of blockchain in healthcare: a systematic literature review. Int. J. Adv. Comput. Sci. Appl. **10**(5) (2019)
11. McGhin, T., Raymond Choo, K.-K., Liu, C.Z., He, D.: Blockchain in healthcare applications: research challenges and opportunities. J. Netw. Comput. Appl. **135**, 62–75 (2019)
12. Tanwar, Sudeep, Parekh, Karan, Evans, Richard: Blockchain-based electronic healthcare record system for healthcare 4.0 applications. J. Inf. Secur. Appl. **50**(102407), 102407 (2020)
13. Hasselgren, A., Kralevska, K., Gligoroski, D., Pedersen, S.A., Faxvaag, A.: Blockchain in healthcare and health sciences-a scoping review. Int. J. Med. Inform. **134**(104040), 104040 (2020)

k-Nearest Neighbor Learning for Secure Intelligent Reflecting Surface Design

Yumou Chen, Muhammad R. A. Khandaker, Sami Azam, Faisal Tariq, and Risala T. Khan

Abstract Recently, intelligent reflecting surfaces (IRSs) have seen an upsurge of interest due to their ability to make the wireless environment programmable, which has historically been treated as an uncontrollable natural phenomenon. Even passive IRSs consisting of many reflecting units can autonomously adjust the reflection coefficients to alter the phase and amplitude of the incident signals. However, optimal reflection design for large IRSs are deemed to be impractical due to the underlying computational complexity. In this paper, we design IRS-assisted programmable wireless environment for secure communication using deep learning techniques. More specifically, we consider the k-nearest neighbor learning algorithm for significantly reducing the computational complexity in IRS design. Simulation results demonstrate the effectiveness of the proposed deep learning-based solutions as compared with traditional alternating optimization.

Keywords Machine learning · 6G · Secrecy rate · DNN · KNN

1 Introduction

In the sixth generation (6G) and beyond communication systems, physical layer wireless signal propagation has been envisioned to be programmable, which had histori-

Y. Chen · M. R. A. Khandaker · S. Azam (✉)
School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK
e-mail: yc38@hw.ac.uk

M. R. A. Khandaker
e-mail: m.khandaker@hw.ac.uk

F. Tariq
James Watt School of Engineering, University of Glasgow, University Avenue, Glasgow, G12 8QQ, UK
e-mail: Faisal.Tariq@glasgow.ac.uk

R. T. Khan (✉)
Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, 1342, Bangladesh
e-mail: risala@juniv.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_15

cally been considered as a natural and uncontrollable phenomenon [1]. Metamaterial-based intelligent reflecting surface (IRS) consists a large number of passive reflective elements. It can reconfigure the radio environment by intelligently adjusting phase angle in the reflected version of the incident signal. Metasurface-based IRS is being considered as a key enabling technology for 6G [1, 2]. In stark contrast to existing transceiver wireless link adaptation technologies, IRS actively modifies the communication channel between transmitter and receiver through its programmable reflecting units. This provides a freedom for further improving the performance of the wireless link and the realization of an intelligent and programmable radio environment [3].

Compared with traditional communication channels, IRS has many significant advantages. Firstly, IRS transmits the signal mainly by passive reflection so it does not need to consume energy. Secondly, its programmability provides more flexibility for signal transmission. Thirdly, it is composed of low-cost material so that it can be easily used on a large scale. Last but not least, the characteristics of the IRS brings enormous prospects for physical layer security to be implemented in practice. In this paper, we leverage this idea to improve communication security through the joint optimization of beamforming at the transmitting access point (AP) and adjust the reflecting coefficient of the IRS. If the reflection units of intelligent reflective surface can be designed appropriately, IRS can be used to reflect a gain signal or a break signal onto a signal to the user and a signal to the eavesdropper, respectively [4]. This will enhance the signal of the user and weakens the signal of the eavesdropper so as to improve the security.

Several works have recently considered IRS-based designs for secure communications [5–7]. While these works demonstrate significant improvement in terms of spectral and energy efficiency, most of the proposed approaches in the literature work in iterative fashion, which is heavily computation demanding. While IRSs are generally meant to be implemented in large scale, the optimal design of such IRS reflections is merely impractical. In addition, the metasurface-based IRSs are intelligent in just name unless they can effectively learn from the environment for optimal reflections. On the other hand, 6G communication systems where IRS is expected to be a core component, will require almost zero latency for end-to-end communications. Therefore, practically implementable IRS designs are essential in order to yield the maximum benefit of the technology. Toward this end, we propose deep learning-based IRS designs for secure wireless communication. More specifically, we consider the k-nearest neighbor learning algorithm for real-time design of the IRS reflections. The objective is to maximize the achievable secrecy rate under transmit power constraint. The proposed approach significantly reduces computational complexity as well as end-to-end communication latency. Simulation results demonstrate that the of the proposed deep learning-based solutions can yield comparable performance with traditional alternating optimization.

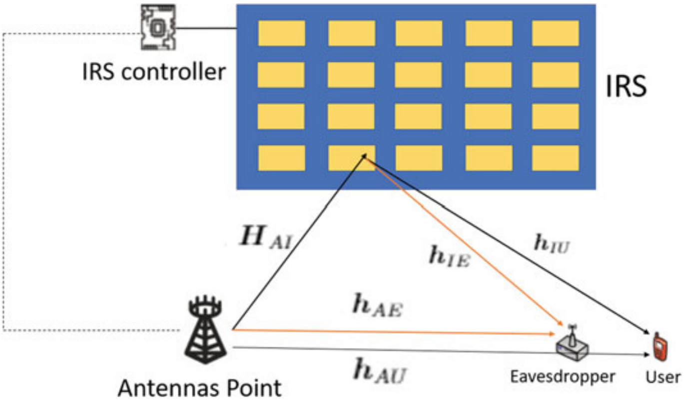


Fig. 1 The model of IRS in the project

2 System Model

Let us consider that an access point (AP) equipped with M antennas wants to transmit a secret message to a single-antenna legitimate receiver with the aid of an IRS with N reflecting units avoiding a single-antenna eavesdropper (Fig. 1). $\mathbf{w}$ is the beamforming vector at the AP. The Euclidean norm of $\mathbf{w}$ needs to satisfy the maximum transmit power constraint at the AP (P_{AP}):

$$||\mathbf{w}||^2 \le P_{\mathrm{AP}}. \tag{1}$$

The reflection units of IRS is denoted by $\mathbf{q} = \left[q[1], q[2], \cdots, q[N]\right]^T$, where $q[n] = \beta_n e^{j\theta_n}$. The phase shift $\theta_n \in [0, 2\pi)$ and the amplitude reflection coefficient $\beta_n \in [0, 2\pi)$, $n = 1, \cdots, N$. In this model, β_n is considered as 1 to get the maximum reflect power. Therefore, $|q[n]| = 1, \forall n$.

In this paper, all the channels are considered as quasi-static flat-fading channels. Let us denote the channels from AP to users, from AP to eavesdropper, from AP to IRS, from IRS to user, from IRS to eavesdropper, respectively, as $\mathbf{h}_{\mathrm{AU}} = \sqrt{\zeta_0 \,(d_0/d_{\mathrm{AU}})^{\alpha_{\mathrm{AU}}}}\mathbf{g}_{\mathrm{AU}}$, $\mathbf{h}_{\mathrm{AE}} = \sqrt{\zeta_0 \,(d_0/d_{\mathrm{AE}})^{\alpha_{\mathrm{AE}}}}\mathbf{g}_{\mathrm{AE}}$, $\mathbf{H}_{\mathrm{AI}} = \sqrt{\zeta_0 \,(d_0/d_{\mathrm{AI}})^{\alpha_{\mathrm{AI}}}}\mathbf{g}_{\mathrm{AI}}$, $\mathbf{h}_{\mathrm{IU}} = \sqrt{\zeta_0 \,(d_0/d_{\mathrm{IU}})^{\alpha_{\mathrm{IU}}}}\mathbf{g}_{\mathrm{IU}}$, $\mathbf{h}_{\mathrm{IE}} = \sqrt{\zeta_0 \,(d_0/d_{\mathrm{IE}})^{\alpha_{\mathrm{IE}}}}\mathbf{g}_{\mathrm{IE}}$, where $\mathbf{g}_{x,y}$ indicates spatially correlated Rician fading model between x and y.

Thus the user's received signal y_U and the eavesdropper's received signal y_E can be respectively expressed as:

$$y_U = (\mathbf{h_{IU}QH_{AI}} + \mathbf{h_{AU}})\mathbf{w} + \mathbf{n_U}, \tag{2}$$

$$y_E = (\mathbf{h_{IE}QH_{AI}} + \mathbf{h_{AE}})\mathbf{w} + \mathbf{n_E}, \tag{3}$$

where $\mathbf{Q}$ is the diagonal matrix with $\mathbf{q}$ as the main diagonal. n_U and n_E denote the Gaussian noises of the user with variances σ_U^2 and that of the eavesdropper with variances σ_E^2 respectively. The user's achievable rate R_U and the eavesdropper's achievable rate R_E can be expressed as:

$$R_U = \log_2\left(1 + \frac{|(\mathbf{h_{IU}QH_{AI}} + \mathbf{h_{AU}})\mathbf{w}|^2}{\sigma_U^2}\right), \tag{4}$$

$$R_E = \log_2\left(1 + \frac{|(\mathbf{h_{IE}QH_{AI}} + \mathbf{h_{AE}})\mathbf{w}|^2}{\sigma_E^2}\right), \tag{5}$$

3 Problem Formulation and Proposed Solution

Our aim is to maximize the communication secrecy rate of the system $R_{\text{sec}} = [R_U - R_E]^+$ by jointly optimizing the transmit beamforming vector from AP and the reflect beamforming vector from IRS. This problem can be formulated as:

$$\begin{aligned}
\max_{\mathbf{w},\mathbf{Q}} \quad & \log_2\left(1 + \frac{|(\mathbf{h_{IU}QH_{AI}} + \mathbf{h_{AU}})\mathbf{w}|^2}{\sigma_U^2}\right) \\
& - \log_2\left(1 + \frac{|(\mathbf{h_{IE}QH_{AI}} + \mathbf{h_{AE}})\mathbf{w}|^2}{\sigma_E^2}\right) && \text{(6a)} \\
\text{s.t.} \quad & ||\mathbf{w}||^2 \leq P_{\text{AP}}, && \text{(6b)} \\
& \mathbf{Q} = \text{diag}\left(\beta_1 e^{j\theta_1}, \cdots, \beta_N e^{j\theta_N}\right), && \text{(6c)} \\
& |q[n]| = 1, \forall n. && \text{(6d)}
\end{aligned}$$

We need to maximize the value of R_{sec} by optimizing $\mathbf{w}$ and $\mathbf{Q}$. The problem (6) is a non-convex problem with a non-convex objective together with non-convex constraint (6d) so that the problem is extremely difficult to solve. In the following, we first apply traditional block coordinate descent (BCD) method, and then apply deep learning techniques to solve the non-convex problem.

3.1 Traditional Iterative Approach

A traditional approach to solving the non-convex problem (6) is to decompose the problem into convex sub-problems and then solve in an alternating fashion [6]. In the alternating optimization algorithm, we first optimize the beamforming vector $\mathbf{w}$ based on given $\mathbf{Q}$, and then optimize $\mathbf{Q}$ with the optimized $\mathbf{w}$. This process is repeated until the required accuracy is reached.

Optimizing w with given Q By denoting

$$\mathbf{A} = \frac{1}{\sigma_{\mathrm{U}}^2} \left(\mathbf{h}_{\mathrm{IU}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} + \mathbf{h}_{\mathrm{AU}}\right)^H \left(\mathbf{h}_{\mathrm{IU}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} + \mathbf{h}_{\mathrm{AU}}\right), \tag{7}$$

$$\mathbf{B} = \frac{1}{\sigma_{\mathrm{E}}^2} \left(\mathbf{h}_{\mathrm{IE}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} + \mathbf{h}_{\mathrm{AE}}\right)^H \left(\mathbf{h}_{\mathrm{IE}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} + \mathbf{h}_{\mathrm{AE}}\right), \tag{8}$$

we obtain from (6)

$$\max_{\mathbf{w}} \quad \frac{\mathbf{w}^H\mathbf{A}\mathbf{w} + 1}{\mathbf{w}^H\mathbf{B}\mathbf{w} + 1} \tag{9a}$$

$$\text{s.t.} \quad \mathbf{w}^H\mathbf{w} \le P_{\mathrm{AP}}. \tag{9b}$$

The optimal solution to problem (9) is given by [8]

$$\mathbf{w}_{\mathrm{opt}} = \sqrt{P_{\mathrm{AP}}}\mathbf{u}_{\max}, \tag{10}$$

where $\mathbf{u}_{\max}$ is the eigenvector corresponding to the largest eigenvalue of the matrix $\left(\mathbf{B} + \frac{1}{P_{\mathrm{AP}}}\mathbf{I}_M\right)^{-1} \left(\mathbf{A} + \frac{1}{P_{\mathrm{AP}}}\mathbf{I}_M\right)$ where $\mathbf{I}_M$ denotes an $M \times M$ identity matrix.

Optimizing q with given w This time we obtain

$$\max_{\mathbf{q}} \quad \frac{\frac{1}{\sigma_{\mathrm{U}}^2}\left|\left(\mathbf{h}_{\mathbf{IU}}\mathbf{Q}\mathbf{H}_{\mathbf{AI}} + \mathbf{h}_{\mathbf{AU}}\right)\mathbf{w}\right|^2 + 1}{\frac{1}{\sigma_{\mathrm{E}}^2}\left|\left(\mathbf{h}_{\mathbf{IE}}\mathbf{Q}\mathbf{H}_{\mathbf{AI}} + \mathbf{h}_{\mathbf{AE}}\right)\mathbf{w}\right|^2 + 1} \tag{11a}$$

$$\text{s.t.} \quad \mathbf{Q} = \mathrm{diag}\left(\beta_1 e^{j\theta_1}, \cdots, \beta_N e^{j\theta_N}\right), \tag{11b}$$

$$|q_n| = 1, \forall n. \tag{11c}$$

Since

$$\begin{aligned} \mathbf{h}_{\mathrm{IU}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} &= \mathbf{q}^T \,\mathrm{diag}\left(\mathbf{h}_{\mathrm{IU}}\right)\mathbf{H}_{\mathrm{AI}}, \\ \mathbf{h}_{\mathrm{IE}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} &= \mathbf{q}^T \,\mathrm{diag}\left(\mathbf{h}_{\mathrm{IE}}\right)\mathbf{H}_{\mathrm{AI}}, \end{aligned} \tag{12}$$

we have

$$\begin{aligned} \tfrac{1}{\sigma_{\mathrm{U}}^2}\left|\left(\mathbf{h}_{\mathrm{IU}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} + \mathbf{h}_{\mathrm{AU}}\right)\mathbf{w}\right|^2 &= \mathbf{s}^H\mathbf{G}_{\mathrm{U}}\mathbf{s} + h_{\mathrm{U}} \\ \tfrac{1}{\sigma_{\mathrm{E}}^2}\left|\left(\mathbf{h}_{\mathrm{IE}}\mathbf{Q}\mathbf{H}_{\mathrm{AI}} + \mathbf{h}_{\mathrm{AE}}\right)\mathbf{w}\right|^2 &= \mathbf{s}^H\mathbf{G}_{\mathrm{E}}\mathbf{s} + h_{\mathrm{E}} \end{aligned} \tag{13}$$

where $\mathbf{s} = \left[\mathbf{q}^T, 1\right]^T$, $\mathbf{G}_{\mathrm{U}}$, $\mathbf{G}_{\mathrm{E}}$ are defined at the top of the next page, and

$$h_{\mathrm{U}} = \frac{\mathbf{h}_{\mathrm{AU}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{h}_{\mathrm{AU}}^T}{\sigma_{\mathrm{U}}^2}, h_{\mathrm{E}} = \frac{\mathbf{h}_{\mathrm{AE}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{h}_{\mathrm{AE}}^T}{\sigma_{\mathrm{E}}^2}. \tag{14}$$

$$\mathbf{G}_{\mathrm{U}} = \frac{1}{\sigma_{\mathrm{U}}^2}\begin{bmatrix} \operatorname{diag}\left(\mathbf{h}_{\mathrm{II}}^*\right)\mathbf{H}_{\mathrm{AI}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{H}_{\mathrm{AI}}^T\operatorname{diag}\left(\mathbf{h}_{\mathrm{IU}}\right) & \operatorname{diag}\left(\mathbf{h}_{\mathrm{IU}}^*\right)\mathbf{H}_{\mathrm{AI}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{h}_{\mathrm{AU}}^T \\ \mathbf{h}_{\mathrm{AU}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{H}_{\mathrm{AI}}^T\operatorname{diag}\left(\mathbf{h}_{\mathrm{IU}}\right) & 0 \end{bmatrix} \tag{15}$$

$$\mathbf{G}_{\mathrm{E}} = \frac{1}{\sigma_{\mathrm{E}}^2}\begin{bmatrix} \operatorname{diag}\left(\mathbf{h}_{\mathrm{IE}}^*\right)\mathbf{H}_{\mathrm{AI}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{H}_{\mathrm{AI}}^T\operatorname{diag}\left(\mathbf{h}_{\mathrm{IE}}\right) & \operatorname{diag}\left(\mathbf{h}_{\mathrm{IE}}^*\right)\mathbf{H}_{\mathrm{AI}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{h}_{\mathrm{AE}}^T \\ \mathbf{h}_{\mathrm{AE}}^*\mathbf{w}^*\mathbf{w}^T\mathbf{H}_{\mathrm{AI}}^T\operatorname{diag}\left(\mathbf{h}_{\mathrm{IE}}\right) & 0 \end{bmatrix} \tag{16}$$

By substituting (13) and into (11), the problem can be simplified to

$$\max_{\mathbf{s}} \quad \frac{\mathbf{s}^H\mathbf{G}_{\mathbf{U}}\mathbf{S} + h_{\mathbf{U}} + 1}{\mathbf{s}^H\mathbf{G}_{\mathbf{E}}\mathbf{S} + h_{\mathbf{E}} + 1} \tag{17a}$$

$$\text{s.t.} \quad \mathbf{Q} = \operatorname{diag}\left(\beta_1 e^{j\theta_1}, \cdots, \beta_N e^{j\theta_N}\right), \tag{17b}$$

$$\mathbf{s}^H\mathbf{E}_n\mathbf{s} = 1, \forall n \tag{17c}$$

where the (i, j)th element of $\mathbf{E}_n$ satisfies

$$[\mathbf{E}_n]_{i,j} = \begin{cases} 1 & i = j = n \\ 0 & \text{otherwise}. \end{cases} \tag{18}$$

By applying semi-definite relaxation (SDR) technique [9], and denoting $\mathbf{S} \triangleq \mathbf{s}\mathbf{s}^H$, the problem (17) can be converted to a relaxed form

$$\max_{\mathbf{S}\succeq\mathbf{0}} \quad \frac{\operatorname{tr}\left(\mathbf{G}_{\mathrm{U}}\mathbf{S}\right) + h_{\mathrm{U}} + 1}{\operatorname{tr}\left(\mathbf{G}_{\mathrm{E}}\mathbf{S}\right) + h_{\mathrm{E}} + 1} \tag{19a}$$

$$\text{s.t.} \quad \operatorname{tr}\left(\mathbf{E}_n\mathbf{S}\right) = 1, \forall n \tag{19b}$$

Next, setting $\mu = 1/\left[\operatorname{tr}\left(\mathbf{G}_{\mathrm{E}}\mathbf{S}\right) + h_{\mathrm{E}} + 1\right]$ and $\mathbf{X} = \mu\mathbf{S}$ and applying Charnes Cooper transformation [9], the problem can be expressed in non-fractional form

$$\max_{\mu\geq 0, \mathbf{X}\succeq\mathbf{0}} \quad \operatorname{tr}\left(\mathbf{G}_{\mathbf{U}}\mathbf{X}\right) + \mu\left(h_{\mathbf{U}} + 1\right) \tag{20a}$$

$$\text{s.t.} \quad \operatorname{tr}\left(\mathbf{G}_{\mathrm{E}}\mathbf{X}\right) + \mu\left(h_{\mathrm{E}} + 1\right) = 1 \tag{20b}$$

$$\operatorname{tr}\left(\mathbf{E}_n\mathbf{X}\right) = \mu, \forall n. \tag{20c}$$

Problem (20) is a convex semi-definite programming (SDP) problem, which can be solved by interior point method [10]. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Conventional iterative algorithm

Initialize: Set $k = 0$; $\mathbf{w}^{[\mathbf{0}]} = \frac{\mathbf{h}_{\mathbf{AI}}^{\mathbf{H}}}{||\mathbf{h}_{\mathbf{AI}}||}$; $\mathbf{q}^{(\mathbf{0})} = \mathbf{1_N}$ and $R^{(0)} = f(\mathbf{w}^{(\mathbf{0})}, \mathbf{q}^{(\mathbf{0})})$.
repeat
Set $k = k + 1$;
With given $\mathbf{q}^{(k-1)}$, find the $\mathbf{u}_{\max}$ of $\left(\mathbf{B} + \frac{1}{P_{\mathrm{AP}}}\mathbf{I}_M\right)^{-1}\left(\mathbf{A} + \frac{1}{P_{\mathrm{AP}}}\mathbf{I}_M\right)$. Set $\mathbf{w}^{(k)} = \sqrt{P_{\mathrm{AP}}}\mathbf{u}_{\max}$.
With given $\mathbf{w}^{(k)}$, solve problem (20) to get the approximate solution $\mathbf{q}^{(k)}$ after standard Gaussian randomization.
Set $R^{(k)} = f(\mathbf{w}^{(\mathbf{k})}, \mathbf{q}^{(\mathbf{k})})$.
until $\frac{R^{(k)} - R^{(k-1)}}{R^{(k)}} < \epsilon$.

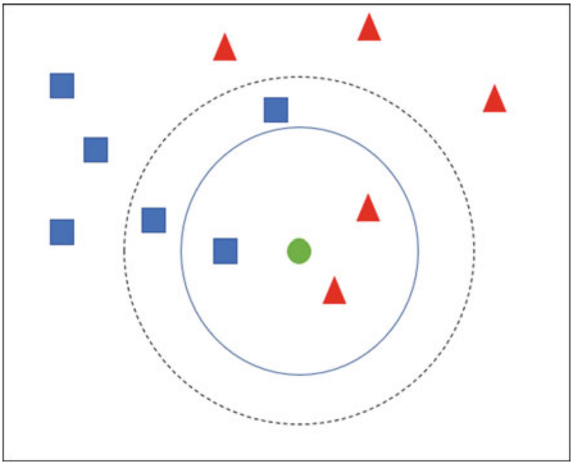


Fig. 2 Principle of judgment in kNN algorithm

3.2 Proposed k-Nearest Neighbor Algorithm

The *k*-Nearest Neighbor (kNN) classification algorithm (aka proximity algorithm), is one of the simplest classification algorithms using deep learning. kNN algorithm classifies by measuring the distance between different feature values.

To be more specific, the sample will be divided into a class if most of the *k* nearest samples of the current sample belong to the certain class. In the kNN algorithm, all comparison samples are correctly classified in advance. This algorithm classifies by comparing all the samples to get the samples with the most similar features.

For example, if the green dot in Fig. 2 represents an unknown category, the kNN algorithm needs to classify it as belonging to a red triangle or a blue square. First, the kNN algorithm needs to get the value of k (how many nearest samples are taken as classification criteria). When k is equal to 3, the closest three colors are red, red, and blue, so the kNN algorithm classifies the dot as red class. If k is equal to 5, the green dot is classified to a blue class because the blue square ratio is 3/5.

Since kNN algorithm is used to solve classification problem, we need to transmit the regression problem to classification problem. In problem (6), we set a million

Algorithm 2 Proposed kNN Algorithm

1: Calculate the distance between the point in the data set of a known category and the current point;
2: Sort in increasing distance order;
3: Select the k points with the smallest distance from the current point;
4: Determine the frequency of the first k points in the category;
5: Return the category with the highest frequency of the first k points as the predicted classification of the current point.

random data sets of channels and divide the **q** and **w** into ten equal parts within their value ranges. Then find a set **q** and **w** that maximizes secrecy rate as the class of the channel. After a sufficiently large training set is established in this way, it just need to find the most similar channels after comparison when entering new channel parameters and get the value of **q** and **w** according to its class in common kNN algorithm.

3.3 Deep Supervised Learning Approach

The neural network layer constructed by the model in this paper is the dense neural network layer or called the fully connected layer which means all the upper layer and the next layer of neurons are fully connected. Such a dense network connection can better reflect the real human brain connection and make full use of all features for training.

After the data pre-processing generates features and label, they need to be input into the multi-layer perceptron model for training to establish a deep learning training model. The flowchart of DNN model is shown in Fig. 3, we can see that it updates hidden layers by following the result of calculated error.

After building a deep learning model, a back propagation DNN algorithm is used for training. Back propagation algorithms are often used in conjunction with optimizers to train multi-layer artificial neural networks [11, 12]. The back propagation algorithm is a supervised learning method which needs to input features and labels. By continuously learning the weights and biases of the DNN, an optimizer is used to minimize the difference between the output value of the deep neural network and the true target value. In this paper, we choose to use stochastic gradient descent (SGD) algorithm. Here are the steps,

Algorithm 3 Supervised Learning Algorithm

1: Establish a multi-layer perceptron model, and initialize the weights and bias of the model with random numbers.
2: Input the features to the neural network for calculation.
3: Calculate the output result through multiple hidden layer networks.
4: Use the loss function to calculate the error between the output and the label. In this project, the loss function uses the cross entropy function [13], which can be expressed as:

$$\delta^L = \frac{\partial f(W,b)}{\partial z^L} = \frac{\partial f(W,b)}{\partial a^L} \odot \sigma'\left(z^L\right) \tag{21}$$

where $\odot$ is the circle convolution or called cyclic convolution, W is the weights, b is the bias, δ^L is the error value of output layer, $z^{\mathbf{L}}$ is the neuron input of output layer, a^L is the neuron output of output layer and σ is the activation function.
5: According to the error value, the weights and deviations of neuron connections are reversely updated layer by layer. According to mathematical induction,

$$\delta^l = \delta^{l+1}\frac{\partial z^{l+1}}{\partial z^l} = \left(W^{l+1}\right)^T \delta^{l+1} \odot \sigma'\left(z^l\right) \tag{22}$$

6: Then the gradient expression is

$$\frac{\partial f(W,b)}{\partial W'} = \frac{\partial f(W,b,x,y)}{\partial z'}\frac{\partial z'}{\partial W'} = \delta^l\left(a^{l-1}\right)^T \tag{23}$$

7: Further, we can use the gradient descent method to optimize the model until we get the required accuracy.

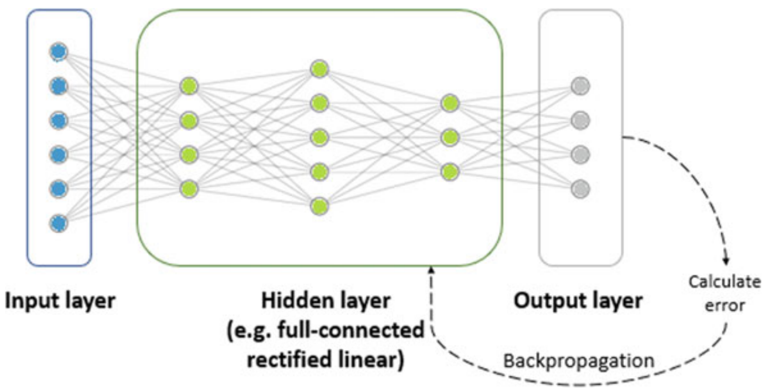


Fig. 3 The model of DNN

Table 1 Simulation parameters

Simulation parameter	Symbol	Value
AP location	$\mathbf{l_{AP}}$	$[0, 0]$
User location	$\mathbf{l_U}$	$[150, 0]$
Eavesdropper location	$\mathbf{l_E}$	$[145, 0]$
IRS location	$\mathbf{l_{IRS}}$	$[145, 5]$
Path loss	ζ_0	-30 dB
The reference distance	d_0	1 m
The distance from AP to user	d_{AU}	150 m
The distance from AP to eavesdropper	d_{AE}	145 m
The distance from AP to IRS	d_{AI}	145.086 m
The distance from IRS to user	d_{IU}	7.071 m
The distance from IRS to eavesdropper	d_{IE}	5 m
The path loss exponents between AP and user	α_{AU}	3
The path loss exponents between AP and eavesdropper	α_{AE}	3
The path loss exponents between AP and IRS	α_{AI}	2.2
The path loss exponents between IRS and user	α_{IU}	3
The path loss exponents between IRS and eavesdropper	α_{IE}	3
Rician factor	K	1
The spatial correlation matrix	$\mathbf{R}$	$[\mathbf{R}]_{i,j} = 0.95^{\lvert i-j \rvert}$
Variance of Gaussian noise at user	σ_U^2	-80 dBm
Variance of Gaussian noise at eavesdropper	σ_E^2	-80 dBm
Accuracy	ϵ	10^{-3}

4 Simulation Results

In this section, we evaluate the performance of the proposed approach. In all the simulations, we used the parameters as described in Table 1 unless otherwise specified. We chose the spatially correlated Rician fading model.

Figure 4 shows the system secrecy rates versus the transmit power with different algorithms. It is clear that optimizing AP alone without using IRS ($\mathbf{q} = 0$) performs the worst. The reason is that the communication channel used by the user is highly correlated with the communication channel used by the eavesdropper. As for other algorithms which use IRS, IRS can adjust the reflecting coefficients of reflecting units to add gain signal to the signal of users to strengthen signal power.

Compared with the beamforming without IRS, it is obvious that IRS can improve a lot in communication secrecy. The AP MRT line denotes the method that uses maximum ratio transmission (MRT)-based beamforming at the channel between the IRS and the AP. The performance of alternating optimization (proposed in [6]) result is better than the performance of AP MRT with IRS result because the alternating optimizing algorithm jointly optimizes both IRS reflect beamforming and AP

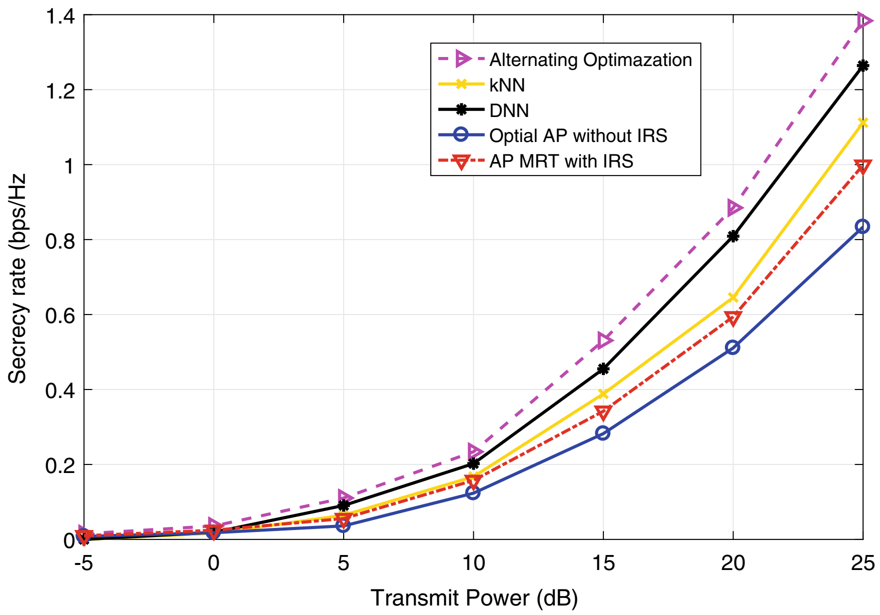


Fig. 4 Secrecy rate versus transmit power

transmit beamforming. However, the computational complexity of the 'Alternating Optimization' is significantly higher.

We also apply two machine learning algorithms to compare with traditional alternating algorithms. It is obvious that the optimization performance of these two ML algorithms is comparable to that of traditional algorithms. The DNN (supervised learning) method performs much closer to the 'Alternating Optimization'. The reason is that the training set is taken from traditional algorithm in [6]. For the kNN line, it has the theoretical possibility to surpass the traditional algorithm. Its training set verification method is similar to the exhaustive method. In Fig. 5, we can observe how the error rate of kNN decreases with the increasing of training set size. If the training set can be raised to a sufficiently large level, the kNN algorithm can theoretically outperform the traditional algorithm.

Although the performance of the two machine learning algorithms is not as good as the traditional algorithms, they have their own advantages. Traditional algorithms require a lot of mathematical calculations after each signal is emitted to IRS and then adjust the IRS reflection coefficient. For ML methods, they have trained the entire model which can get the result of IRS reflection coefficient immediately without many mathematical calculations.

DNN algorithm has better performance than kNN algorithm, but the generation of training sets for kNN are much easier than that of DNN [14, 15]. In Fig. 6, we can observe that the time required for DNN to build a model is much longer than kNN and the gap is still getting larger as the amount of data becomes larger. The

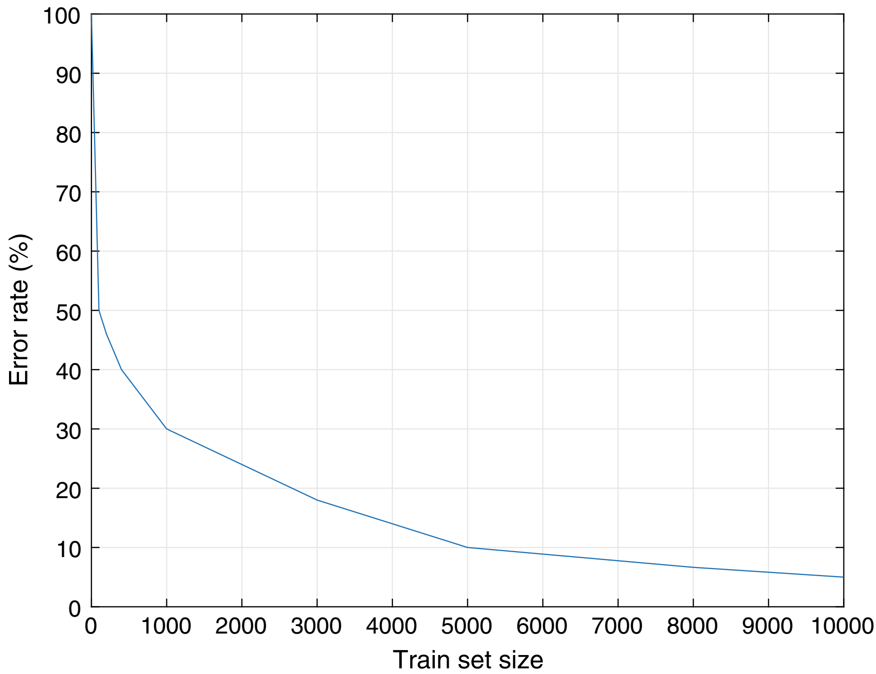


Fig. 5 Error rate of kNN varies with size of the training data set

reason is that the training set of kNN algorithm only needs exhaustive method to get the maximum value while DNN algorithm needs a lot of calculation by traditional algorithm. Therefore, under the premise of large scale application, kNN algorithm has the advantage of saving more cost.

Figure 7 illustrates the average secrecy rates versus the number of reflecting units (N) when the transmit power is 15 dBm. The optimal AP without IRS line does not increase because it does not have IRS. The performance of all the algorithm with IRS becomes better with the increasing of N. As the increasing of size of IRS, the beamforming can be more effective. In addition, the performance of DNN is very close to that of traditional alternating algorithms when the reflecting units of IRS are enough.

By comparing the three solutions, the traditional alternative optimization algorithm has the best performance, and the machine learning algorithm can get the results quickly. In practice, the large amount of computation required by the traditional algorithm cannot meet the requirements of future communication systems. The DNN performs better with the computational power of the personal computer but the kNN has a higher ceiling if the computational power can be increased significantly.

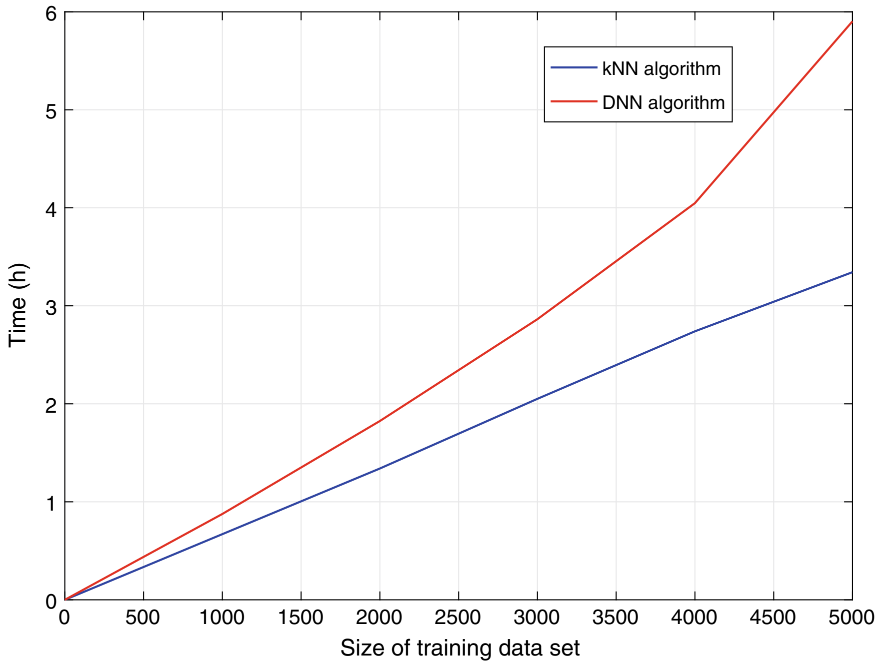


Fig. 6 The time of building a model versus the size of the data

5 Conclusion

The IRS is a cost-effective technology consisting of a large number of low-cost reflection units, which can greatly improve the performance of the physical layer without incurring the high cost and power consumption required for multiple antennas. We have introduced a truly intelligent reflecting surface aided secure communication system. After modeling and formulation of practical problems, the solution of the problem is analyzed by comparing the advantages and disadvantages of traditional algorithm (alternating optimization) and machine learning algorithm (*k*-Nearest Neighbor Algorithm and Deep Neural Networks) are applied. Simulation results demonstrate that the proposed learning approaches can achieve comparable performance with the alternating optimization algorithm, and is simpler to implement, shorter in operation time and offers greater flexibility.

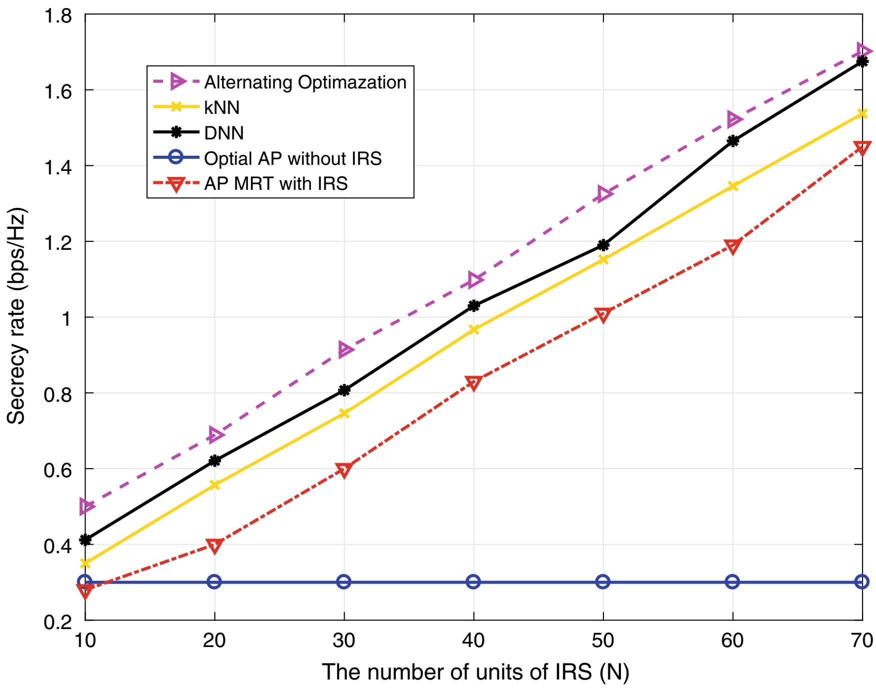


Fig. 7 Secrecy rate versus different number of reflecting surface

References

1. Tariq, F., Khandaker, M.R.A., Wong, K.-K., Imran, M.A., Bennis, M., Debbah, M.: A speculative study on 6G. IEEE Wireless Commun. **27**(4), 118–125 (2020)
2. Zhao, J.: A survey of intelligent reflecting surfaces (IRSS): towards 6g wireless communication networks (2019)
3. Basar, E., Di Renzo, M., De Rosny, J., Debbah, M., Alouini, M., Zhang, R.: Wireless communications through reconfigurable intelligent surfaces. IEEE Access **7**, 116 753–116 773 (2019)
4. Tan, X., Sun, Z., Jornet, J.M., Pados, D.: Increasing indoor spectrum sharing capacity using smart reflect-array (2015)
5. Chen, J., Liang, Pei, Y., Guo, H.: Intelligent reflecting surface: a programmable wireless environment for physical layer security. IEEE Access **7**, 82 599–82 612 (2019)
6. Cui, M., Zhang, G., Zhang, R.: Secure wireless communication via intelligent reflecting surface. IEEE Wireless Commun. Lett. **8**(5), 1410–1414 (2019). Oct
7. Yu, X., Xu, D., Schober, R.: Enabling secure wireless communications via intelligent reflecting surfaces. In: 2019 IEEE Global Communications Conference (GLOBECOM) (2019)
8. Khisti, A., Wornell, G.W.: Secure transmission with multiple antennas i: the misome wiretap channel. IEEE Trans. Inf. Theo. **56**(7), 3088–3104 (2010)
9. Liu, L., Zhang, R., Chua, K.: Secrecy wireless information and power transfer with miso beamforming. IEEE Trans. Sig. Proc. **62**(7), 1850–1863 (2014)
10. Wright, S.J.: Effects of finite-precision arithmetic on interior-point methods for nonlinear programming. SIAM J. Optim. **12**(1), 36–43 (2001)

11. Várkonyi-Kóczy, A.R., Tusor, B.: Improved back-propagation algorithm for neural network training. In: 2011 IEEE 7th International Symposium on Intelligent Signal Processing, pp. 1–8 (2011)
12. Zhixin, S., Bingqing, L.: Research of improved back-propagation neural network algorithm. In: 2010 IEEE 12th International Conference on Communication Technology, pp. 763–766 (2010)
13. Zhou, Y., Wang, X., Zhang, M., Zhu, J., Zheng, R., Wu, Q.: Mpce: a maximum probability based cross entropy loss function for neural network classification. IEEE Access **7**, 146 331–146 341 (2019)
14. Mahmud, M., Kaiser, M.S., Hussain, A., Vassanelli, S.: Applications of deep learning and reinforcement learning to biological data. IEEE Trans. Neural Netw. Learn. Syst. **29**(6), 2063–2079 (2018)
15. Mahmud, M., Kaiser, M.S., McGinnity, T.M., Hussain, A.: Deep learning in mining biological data. Cognit. Comput. **13**(1), 1–33 (2021)

Smart Home Surveillance Based on IoT

Lutfun Nahar, Md. Shahadat Hossain, Nusrat Jahan, Mayesha Tasnim, Karl Andersson, and Md. Sazzad Hossain

Abstract Internet of things (IoT) is a process of inter-networking of constitutional design, vehicles (also referred to as "connected devices" and "smart devices"), and various objects—included with software, electronics, actuators, sensors, and inter-network connectivity which is used to gather and interchange of data and information. IoT can control objective and gather information from distance across the network. Recently smart home systems attained more popularity to increase the quality of life where smartphone application plays an important role to monitor home apparatus using wireless communication. The target of this research is to develop an IoT-based smart home to control specific devices and monitoring house using android mobile devices. Features that are used here are temperature, gas detection, and door lock. A channel, alarm system, door lock can be visualized by the users from any mobile device. If any change happened in the monitored data, user will be notified.

Keywords IoT · Smart home · Temperature · Gas · Door lock · Smart bin

1 Introduction

IoT is a process of interconnected counting instrument, digital instruments, object, animals, and people that are giving the power to change data over an interconnection without demanding interaction between human and computer. In IoT, Internet with smart technology is used to establish communication between devices using

L. Nahar · N. Jahan · M. Tasnim
International Islamic University Chittagong, Chittagong, Bangladesh

Md. S. Hossain (✉)
University of Chittagong, Chittagong, Bangladesh
e-mail: hossain_ms@cu.ac.bd

K. Andersson
Lulea University of Technology, Skelleftea, Sweden

Md. S. Hossain
University of Liberal Arts Bangladesh, Dhaka 1209, Bangladesh
e-mail: sazzad.hossain@ulab.edu.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_16

actuators, and wireless sensors that has been controlled remotely and also monitored by the people for automation is mentioned as Internet of things. The term "Things" widely used for physical instruments like sensors, microcontroller, cameras, and other devices.

A smart home is a concept that uses devices which connects through Internet to enable the remote monitoring and management. It associates everything in the network by Internet. In this term, we can monitor homey apparatus like fluorescence, air cooling, ventilator, TV, fridge, automatic washer, safety technique, etc. It can also be self-discipline and advised itself by using artificial intelligence. These cases additionally aid improvement home's vigor usefulness, which can minimize vigor bill.

The basic most essentials objectives of home automation are to save users time and making livelihood common and standard. This research focuses on implementation of smart house that can monitor house air condition by pressing a key. On the other hand, there is an Android app for controlling the home apparatus.

This research integrates all features of home apparatus so that one can easily operate their home apparatus from distance by using smart device.

The remaining part of the paper is prepared as follows. Relevant works on smart home automation system describe in Sect. 2. Section 3 outlines the system description. Section 4 shows the results, and comparison is illustrated in Sect. 5, while the Sect. 6 gives the conclusion and discussion on further research.

2 Literature Review

There are several works related to IoT. In this paper [1], they gathered a lot of knowledge about IoT. They implemented keyless entry and motion detection and smartphone alert system.

In [2], they used Raspberry Pi and smartphone application users to control a series of home appliances and sensor.

In [3], the planned method not only instruct the visual data, like thermal reading, oratory, fluorescence, displacement transducers, but also incites a method conforming to the necessity of light on if it gets gloomy. It also saves the sensor limitations in the cloud (Gmail) in a sharp method. This co-operated the user to explore the situation of varied restrictions in the house at any moment.

In [4–6], IoT home automation is introduced where an ongoing home automation apparatus with short and obvious view was focused. The benefits on home automation, such as alleviate installation value, system durability, simple augmentation, esthetical advantages, integration of mobiles devices are practiced here.

In [7], they implement a prototype for house automation at peak hour where they use WSN devices. The current research of USA on wireless sensor, integration of artificial intelligence and IoT, nanotechnology, and RFID provides us the idea of RFID and IoT. This research focused on implementation of intelligence office and house system for the foreigners which give automatic information time to time. Household, medical, and environmental appliances are also declared here.

In [8–11], new age house automation is elaborately discussed. In this method, they permitted the user to monitor apparatus and radiances in their house from a cell phone and multiprocessor from any place in the society by a cyberspace relation. It also permitted the user to monitor their instruments systems within their house from house server using GUI. The home server GUI will rule over the method if neither the cell phone nor multiprocessor is capable to rule the device entities in the house.

This is huge informative and work worthy paper based on IoT smart home [12]. In this assessment paper, they explored and appreciated context-aware computing investigation endeavors to feel how the protest in the plot of context-aware computing has been intercept in personal computer, web, cellular phone, device webs, and diffusive computing instance. A huge number of authorizations subsist in terms of methods, multivendor, petitions, technic, and examples proposed by searcher to resolve various protests in context-aware computing. Some of the directions in the plot that recognize by the audit. The consequences fairly reveal the worth of factors sensation in the IoT instance.

In this study [13], a narrative planning for small tariff and gentle house monitor and controlling method using Android formed cellular phone is mentioned and executed. They implemented a web services for transmitting data from distance. Any Android cellular mobile with Wi-Fi can monitor the gadgets of house. In addition without Wi-Fi, cellular phone webbing 4G can also permitted to go into the structure. Further research will give attention on implementing a wireless network using ZIGBEE.

In [14–17], IP-based wireless sensor networks are used to take decision and IP-based WSN is used for smart irrigation system, which predict natural disaster and water quality monitoring [18–21].

In this study [22], IoT-based system is developed for agriculture. In [23, 24], author implemented agro-sensor communication smart system and i-voting smart system. So there is a clear indication that IoT-based smart system plays an important role in our ever day life to make life more easier and comfortable and save time.

Therefore, in this research, IoT-based smart home system is implemented where dock lock, gas detection, and smart been system are introduced. Our proposed system architecture will be described in the following section.

3 System Description

3.1 System Architecture

Smart home defines a system of network, governable device which can work altogether to make our home more relax, customized, proficient, and invulnerable. It makes our life simple. In view of this system come into being five major portions

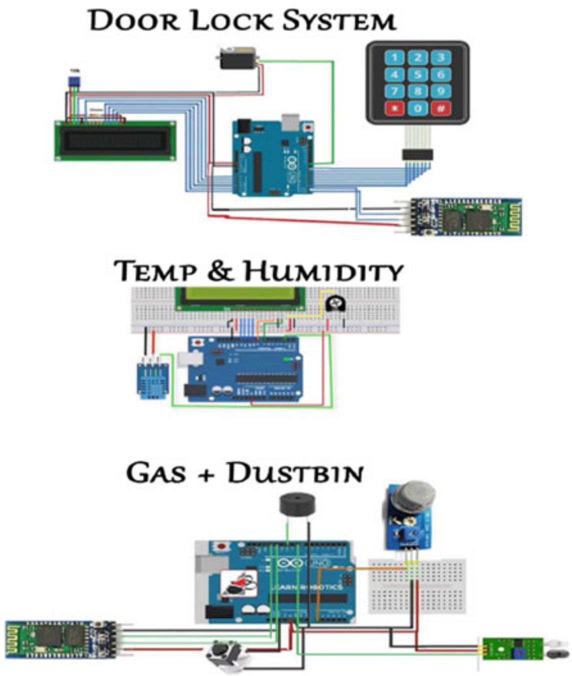


Fig. 1 System architecture

Arduino, relay drivers, Bluetooth module, Android application, and step down transformer. At first, we have to give power to the step down transformer, in step down the input voltage and provide to the Arduino with VIN pin. The Bluetooth module is also need to associate with Arduino to Rx and Tx pin that gives the news to the microcontroller. Microcontroller reads the facts and dispatch to transfer the drivers which exertion as control. In Arduino, we transmit the program as per necessity, then it carries some mathematical and rational functioning to monitor the relay drivers.

System architecture is given in Fig. 1.

3.2 Block Diagram

Here is the block diagram of this system where we showed the overall process of our work.

3.3 Hardware

Arduino Uno: It is a microcontroller fume depending on the Atmega328 (datasheet) with 14 I/O pins, in which six pins can be turned to as yields, six pins are set as simple information sources. It has 16 MHz clay resonator, a USB association, a power jack, and a reset button. The microcontroller has 32 kB of ISP flash memory, 2 kB RAM, and 1 kB EEPROM. The board provides serial indication competence via UART, SPI, and 12C. Cause of well scheme in the structure of Arduino it is simple to feel. In Arduino, we conduct high level of programming language like C language, C++ language, etc. It is simple to feel and user amicable language. It has exceedingly precedence like multitasking, mechanization, time domain, etc. Arduino Uno is given below.

Bluetooth Module: We are using an Android app to connect the project and control it for remote connection between app and Arduino we are using Bluetooth module (master or slave at a time). HC-05 Bluetooth module is conducted to an Android application with the microcontroller. Bluetooth sends to the microcontroller and receives the information from user (Arduino). It is easy to conduct Bluetooth Serial Port Protocol (SSP), planned as transmission sequential connection established. The Bluetooth of serial port module is Advanced Bluetooth v2.0+ Enhanced data Rate at 3Mbps modulation with 2.4 GHz radio receiver with base band (BB). The Bluetooth of Rx and Tx pins are joined to the Arduino pins of Tx and Rx in some respect. To utilize Bluetooth Serial Port Protocol (SPP) module HC-05 module is simple and resolve for directly distant pursuant association setup. It puts CSR Blue canter 04-External single fume Bluetooth framework with CMOS innovation and with Adaptive Frequency Hopping Feature (AFH).

IR Sensor: It is an electronic instrument, which can send in apropos sense various perspectives of the adjacent. An IR sensor can magnitude the warmth of an device like finds the motion. Only infrared radiation can control these type of radiation, more than transmitting it especially called a passive IR sensor. Habitually, in the electromagnetic spectrum, all the purpose diffuse several structure of thermal radiation. We use IR sensor in our project in dustbin module. The dustbin module in our project sense the objects located in front of it. Here, IR sensor is used to sense any obstacle and object; thus, we come to know that the dustbin is full.

Potentiometer: It is a three-terminal instrument panel with a turning contact or sliding which can set up a constant voltage divider. If only two terminals are used, one end and the wiper, it acts as a changeable resistor. In our work, we use this to monitor the contrast and brightness of LCD display.

PCB Board: It electrically joins electronic materials using conductive pads, tracks, and other components inscribe from one or more over sheet layers of copper laminated onto or/and among sheet layers of a non-conductive substrate. Elements are usually sell onto the PCB to both mechanically fasten and electrically join them to it. Here in our project, there are lot of cables and circuits to integrate the spread circuits in one circuit and minimize the cable connection we used PCB board.

Gas sensor: A gas sensor is an instrument that can find the concentration or appearance of gases in the environment. Directed on the concentration of the gas sensor creates an analogous potential difference by switching the resistance of the material within the sensor that can be adequate as output voltage. Based on this voltage standard, the type and concentration of the gas can be calculated. Our project also alarm the user if it find any gas around of it. So we used gas sensor to sense the gas.

Buzzer: The buzzer is a sounding instrument which can change audio gesture into sound gestures. It is commonly powered by DC voltage. It is broadly used in dread, computers, printers, and other electronic manufactures as sound instrument. It is basically parted into electromagnetic buzzer and piezoelectric buzzer described by the letter “HA” or “H” in the circuit. According to different uses and designs, the buzzer can transmit different sounds such as alarm, electric bell, music, buzzer, and siren. To create different alarm for different scenario, we used a couple of buzzer in our buzzer.

Servomotor: A servomotor is a linear actuator that allows for linear or velocity, monitor of angular, and stimulation. It forms of a compatible motor coupled to a sensor for position recompose. In our project, we use servomotor to control the automated door. In our project, we used servomotor that rotates around a predetermined angle.

3.4 Software

Here, Arduino IDE is used for implementing this smart home idea. For connecting the feature with the user, we need to implement the module through proper program. We have to set program in various part. We implement door lock system, gas sensor, temperature sensor, and smart dustbin separately.

Figure 2 describes the main components, and Fig. 3 illustrates processing steps of our proposed system.

Arduino, Bluetooth device, sensor, buzzer, PCB board, and Potentiometer are used for the hardware implementation. The collision buttons of home or loads apparatus are interface to 8051 series microcontroller by the command signals dispatch through RF transmitter after encoding order signals. The RF transmitter end is a RF remote which can be used by the user as a distant controller for operating home apparatus. The receiver end structure of the RF receiver circuit which form of a decoder to decode the encoded order signals receive from the transmitter. The decoded signals are brought up to the microcontroller, and the orders are sent to the conduct loads through opto-isolator.

Here, we use an Android APP named Arduino Bluetooth Controller via this APP we can connect with that Bluetooth which we placed in our hardware component. Once our smartphones Bluetooth connect with the Bluetooth module of our smart home, then it can transfer data and can happen transmit between them to run the project properly and to have expected result from the project. This is a one way transmit at a time the module can work as a sender or a receiver.

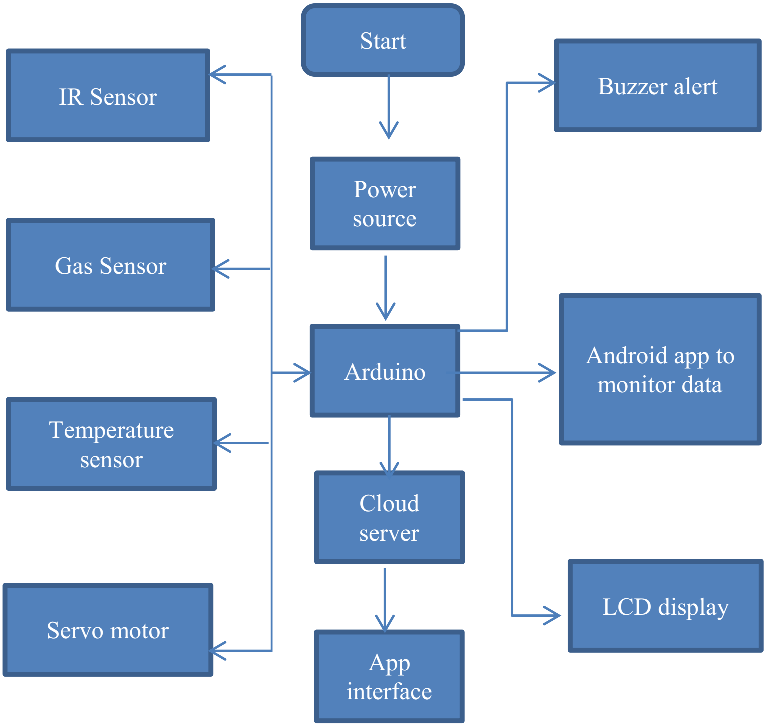


Fig. 2 Block diagram

4 Implementation and Results

A cloud server account has been opened to store information, and then, an API data fetching framework is implemented in the local mobile app system to store data in the cloud. Data are updated time to time in the cloud, and it is monitored from the smart device and also a notification has been send to the mobile device if gas is detected. In addition, an LCD display is added to show bin status, and buzzer alert system is implemented for gas detection. The experimental results for all features have been shown below.

4.1 Door Lock System

The door of the smart home can automatically open or close by servomotor through giving password. Using Arduino control app, we can open or close house door. Figure 4 illustrated the door lock system.

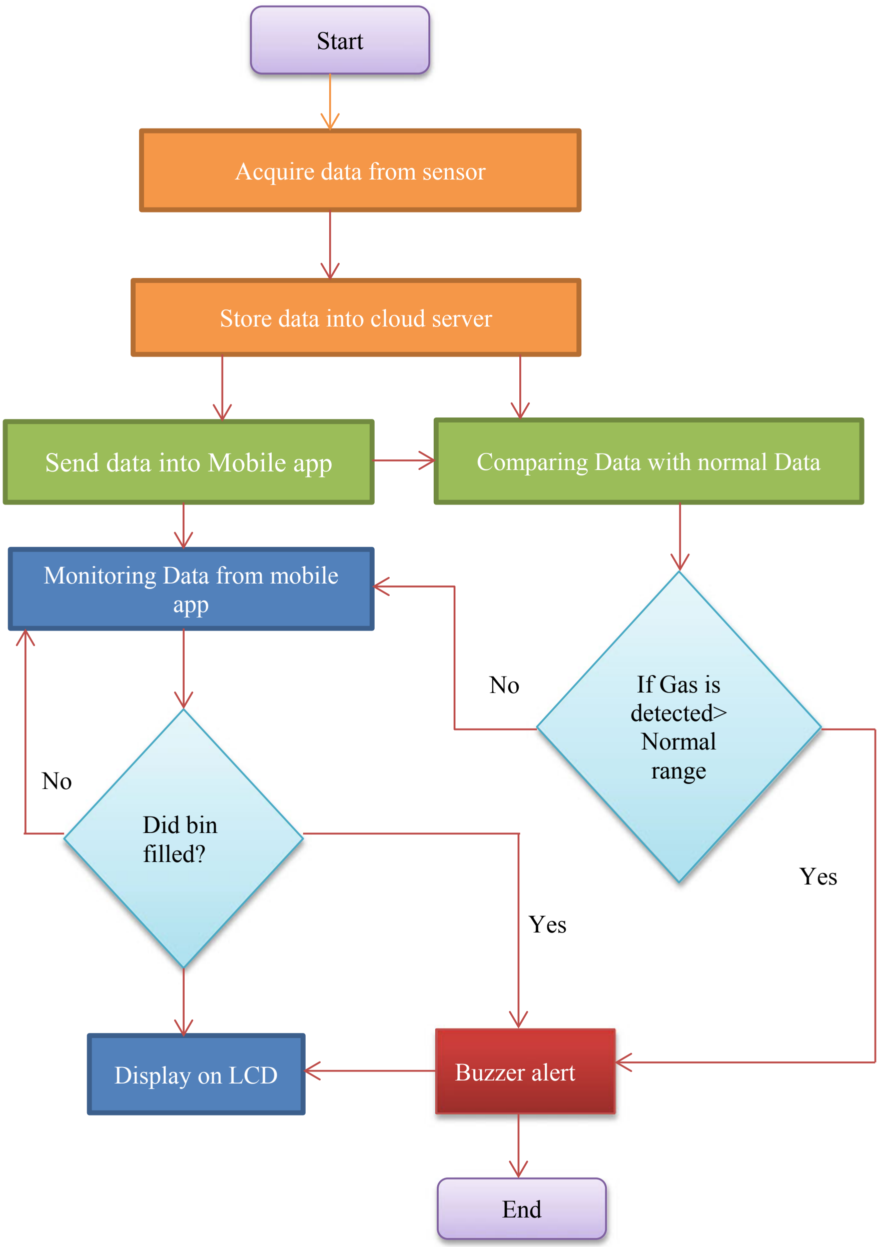


Fig. 3 Work flow diagram

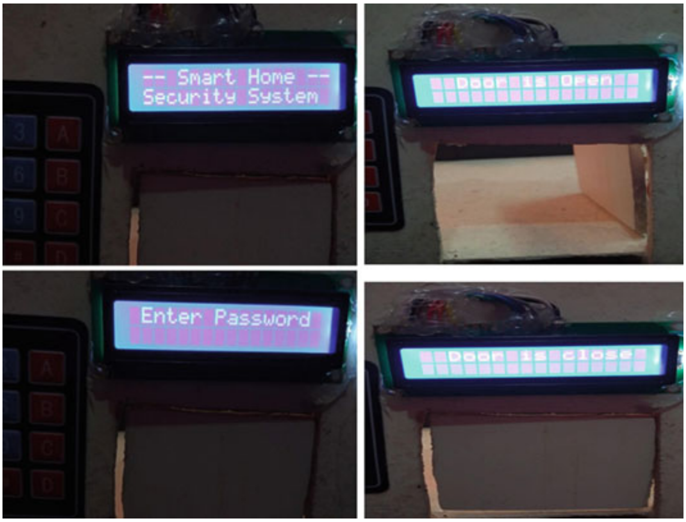
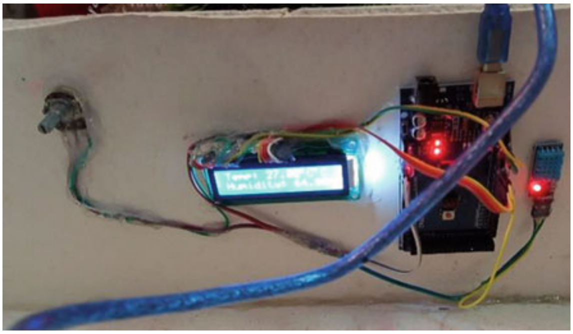


Fig. 4 Door lock system

Fig. 5 Temperature sensor



4.2 Temperature and Humidity System

Dht11 sensor can sense the temperature and humidity and show it in LCD display. Temperature is shown in Fig. 5.

4.3 Gas Detection and Smart Bin System

Smart dustbin is also an element of our smart home which on the buzzer on the user's phone if it can sense any dirt around it. Gas sensor can detect gas around of sensor

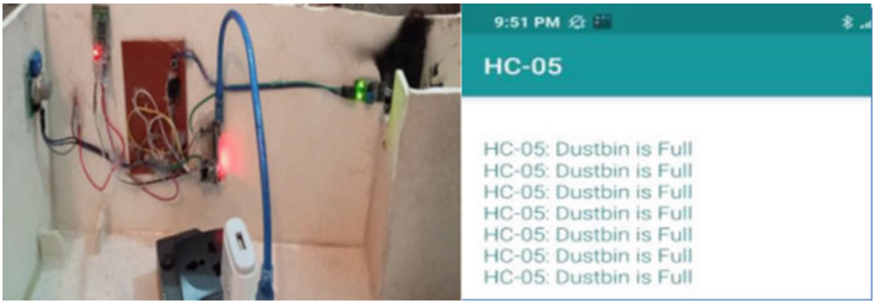


Fig. 6 Gas sensor and smart bin

then giving siren in buzzer. Smart bin alert system and gas detection alert system are shown in Fig. 6.

5 Data Analysis

To validate our proposed system, we compare real time data and system data which is found from Dht11 sensor. The comparisons are shown in Fig. 7.

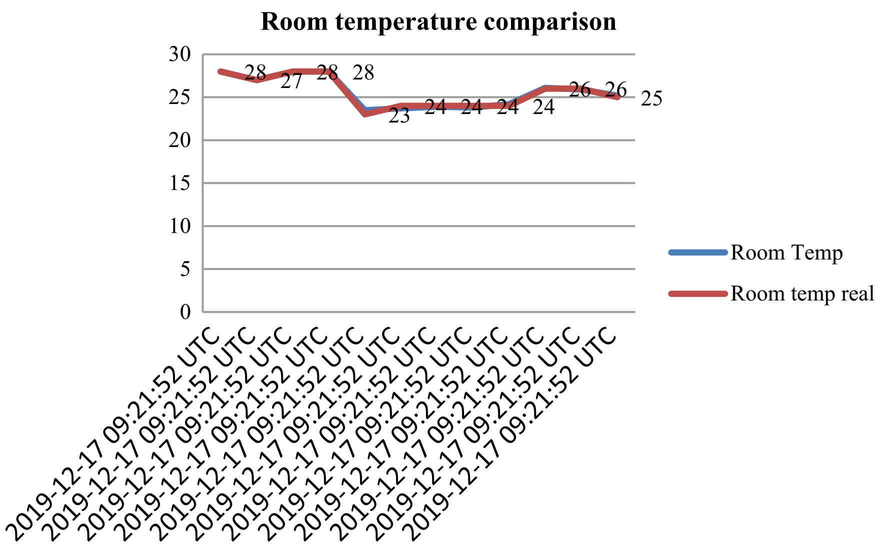


Fig. 7 Comparison of the real data and experimental data

6 Conclusion and Future Work

The goal of this research is to implement a smart home system which can monitor home apparatus with a touch of a finger. An Android app is implemented for controlling the home apparatus. Servomotor is used for door locking system, and Dht11 sensor is used for temperature, humidity, and gas detection. Moreover, an alert system is implemented to detect gas when it is leaked in the house. We make a fully functioning house mechanization technique with facility to keep eye on apparatus and make our life easy and safe.

Further research concern integration of Belief Rule Base Expert system (BRBES) with IoT [25–27]. It is broadly used to handle uncertain knowledge and accomplish the task of evidential reasoning.

References

1. Stolojescu-Crisan, C., Crisan, C., Butunoi, B.: An IoT-based smart home automation system. Sensors **21**, 3784 (2021)
2. Mustafa, B., Iqbal, M.W., Saeed, M., Shafqat, A.R., Sajjad, H., Naqvi, M.R.: IOT based low-cost smart home automation system. In: 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–6, IEEE (2021)
3. Vinay, S.K.N., Kusuma, S.M.: Home automation using Internet of Things. Int. Res. J. Eng. Technol. (IRJET) **02**, 1965–1970 (2015)
4. Li, B., Yu, J.: Research and application on the smart home based on component technologies and Internet of Things. Proc. Eng. **15**, 2087–2092 (2011)
5. Javale, D., Mohsin, M., Nandanwar, S., Shingate, M.: Home automation and security system using android ADK. Int. J. Electronics Commun. Comput. Technol. (IJECCT) **3**, 382–385 (2013)
6. da Cruz, M.A.A., Rodrigues, J.J.P.C., Al-Muhtadi, J., Korotaev, V.V., de Albuquerque, V.H.C.: A reference model for Internet of Things middleware. IEEE Internet Things J **5**, 871–883 (2018)
7. Praveena, B., Rajganesh, K., Vinitha, B., Kandasamy, S., Chandraman, M.: Home security system for controlling theft using android ADK. Int. J. Eng. Res. Technol. **3** (2015)
8. Jain, S., Vinoth Kumar, N., Paventhan, A., Kumar Chinnaiyan, V., Arnachalam, V., Pradish, M.: Survey on smart grid technologies- smart metering, IoT and EMS. In: 2014 IEEE Students Conference on Electrical, Electronics and Computer Science, pp. 1–6. IEEE (2014)
9. Gaikwad, P.P., Gabhane, J.P., Golait, S.S.: A survey based on Smart Homes system using Internet-of-Things. In: 2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), pp. 0330–0335, IEEE (2015)
10. Dickey, N., Banks, D., Sukittanon, S.: Home automation using cloud network and mobile devices. In: 2012 Proceedings of IEEE Southeastcon, pp. 1–4, IEEE (2012)
11. Pavithra, D., Balakrishnan, R.: IoT based monitoring and control system for home automation. In: 2015 Global Conference on Communication Technologies (GCCT), pp. 169–173, IEEE (2015)
12. Perera, C., Zaslavsky, A., Christen, P., Georgakopoulos, D.: Context aware computing for the Internet of Things: a survey. IEEE Commun. Surv. Tutorial **16**, 414–454 (2014)
13. Mohammad El-Basioni, B.M., El-kader, S.M.A., Fakhreldin, M.A.: Smart home design using wireless sensor network and biometric technologies. Int. J. Appl. Innov. Eng. Manage. (IJAIEM) **2**, 413–429 (2013)

14. Thombre, S., Islam, R.U., Andersson, K., Hossain, M.S.: IP based wireless sensor networks: performance analysis using simulations and experiments. J. Wireless Mobile Netw. Ubiquitous Comput. Dependable Appl. **7**(3), 53–76 (2016)
15. Andersson, K., Hossain, M.S.: Heterogeneous wireless sensor networks for flood prediction decision support systems. In: 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (2015)
16. Thombre, S., Ul Islam, R., Andersson, K., Hossain, M.S.: Performance analysis of an IP based protocol stack for WSNs. In: 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (2016)
17. Noor, K., Siddiquee, E., Sarma, D., Nandi, A., Akhter, S., Hossain, S., Andersson, K., Hossain, M.S.: Performance analysis of a surveillance system to detect and track vehicles using Haar cascaded classifiers and optical flow method. In: 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA) (2017)
18. Abedin, Z., Chowdhury, A., Shahadat Hossain, M.S., Andersson, K., Karim, R.: An interoperable IP based WSN for smart irrigation system. In: 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC) (2017)
19. Abedin, Z., Paul, S., Akhter, S., Siddiquee, K., Hossain, M.S., Andersson, K.: Selection of energy efficient routing protocol for irrigation enabled by wireless sensor network. In: 2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops) (2017)
20. Islam, R., Andersson, K., Hossain, M.S.: Heterogeneous wireless sensor networks using CoAP and SMS to predict natural disasters. In: 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (2017)
21. Chowdury, M.S.U., Emran, T.B., Ghosh, S., Pathak, A., Alam, M.M., Absar, N., Andersson, K., Hossain, M.S.: IoT based real-time river water quality monitoring system. Proc. Comput. Sci. **155**, 161–168 (2019)
22. Pathak, A., AmazUddin, M., Abedin, M.J., Andersson, K., Mustafa, R., Hossain, M.S.: IoT based smart system to support agricultural parameters: a case study. Proc. Comput. Sci. **155**, 648–653 (2019)
23. Alam Siddiquee, K., Khan, F., Andersson, K., Hossain, M.S.: Optimal dynamic routing protocols for agro-sensor communication in MANETs. In: 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC) (2017)
24. Alam Siddiquee, K., Andersson, K., Khan, F., Hossain, M.S.: A scalable and secure MANET for an i-voting system. J. Wireless Mobile Netw. Ubiquitous Comput. Dependable Appl. **8**(3), 1–17 (2017)
25. Andersson, K., Hossain, M.S.: Smart risk assessment systems using belief-rule-based DSS and WSN technologies. In: 2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE) (2014)
26. Alam, M., Kaiser, M., Hossain, M.S., Andersson, K.: An IoT-belief rule base smart system to assess autism. In: 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT) (2018)
27. Monrat, A., Ul Islam, R., Hossain, M.S., Andersson, K.: A belief rule based flood risk assessment expert system using real time sensor data streaming. In: 2018 IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops) (2018)

Devices and Communication

Design and Simulation of a Trans-Impedance-Based Instrumental Circuit for Weevil Repelling

M. A. Awal and A. N. M. Ahsan

Abstract A microcontroller-oriented trans-impedance instrumental system was designed and simulated. The system was considered as the alternative to pesticides for repelling weevil from stored rice in Bangladesh. The circuit system was simulated on the open-source TINA-TI SPICE-based analog simulation software. The advantages make the system efficient for repelling weevil. The system has been designed for easy installation, availability at an affordable cost. Effective radiation repelling the weevil ensures grain quality. The system had an application of driving the relevant frequency ultrasonic transducer (30–90 kHz). The radiation was considered with an effective beam angle to the direction of the rice stored area. Relevant low ultrasonic frequency radiation makes the insect embracement. The output signal was calculated as 24 mA and 15 V, and the signal contained 16 harmonics with a total phase change of -18.54 degrees. The system was constructed with the series oscillator and trans-impedance instrumental circuit. This configuration has turned into a quite stable signal output. To achieve the desired low frequency and low distorted ultrasonic signal, a narrow bandpass filter, frequency division, current to voltage conversion, and precision-level pre-amplification instrumentation circuit were introduced in this system. A low harmonic distortion was observed after the simulation of the circuit. The system could be generated any relevant low frequency by uploading a pre-program. The recommended radiation can repel the weevil during the grain infestation. The different low-frequency outputs could also be helpful for other relevant harmful insects.

Keywords RC circuits · DC · DC power circuits · Traditional storage · *Sitophilus Oryzae* · Ultrasonic frequencies · Chemical insecticides · Weevil behavior · Filter · Power amplifier · OPA132 · LM318 · Narrow bandpass filter · Symmetrical output · Trans-impedance · Pre-amps · Acoustic wave

---

M. A. Awal (✉)
Department of Farm Power and Machinery, Bangladesh Agricultural University (BAU), Mymensingh, Bangladesh
e-mail: awalfpm@bau.edu.bd

A. N. M. Ahsan
Precision Lab, Department of Farm Power and Machinery, Bangladesh Agricultural University (BAU), Mymensingh, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_17

1 Introduction

The hot and humid weather of Bangladesh approves the rapid development of the insect population and the deterioration of seeds [1]. The seed's rice grain size and moisture content were identified as the significant reason for weevil infestation. Susceptibility test up to 16th week under the ambient condition (28–32 °C) has identified the highest number of weevil in Nazersail [2]. The weevil eats the endosperm of rice grain which turns the loss of weight and quality of grains. The infestation of rice grains by weevils increases the grain germs and causes poor seed germination [3]. Using practices like phosphine, cowpeas mixed with sieved ashes to limit weevil activity at storage areas are harmful to use and have less effectiveness in reducing insects from the rice storage area [4]. Several technologies are used to control pest insects. Few devices are demonstrated as effective for mosquitoes, cockroaches, and ants [5]. For the alternative of manual and chemical pesticide-related techniques, ultrasound pest repelling devices are introduced in the market. Most of the devices lacked evidence for efficacy, but those continued to sell. Transonic Pro, SonicIQ Ultrasonic pest repeller, model SB105, Pest Free, and Riddex anti-bed bug killer are available in the market. The efficacy test is driven by the influence of radiation of the available devices on bed bugs (*Hemiptera: Cimicidae*). This experiment was suggested that there is a need to develop an effective device for repelling specific insects [6]. Conventional fixed frequency ultrasound equipment may not be suitable for every insect [7]. A particular problem required a different solution. Potential radiation transmission is a useful technology that makes embracement to specific insects. The successful transmission depends on the circuit system design for the high excitation voltages, which count the attenuation due to the propagation of the signal through the air. Usually, the transducer determines the signal transmission and receiving through the same transducer. When this dual operation occurs, it loses the ability to operate the transducer at a serial resonance. The proposed system was considered only the transmission for the selected transducer. The high output voltage and a wide BW-based power amplifier effectively drive the relative transducers [8]. However, it is difficult to find the exact frequency-based crystal oscillator where 80–90 kHz crystals are pretty unavailable. Nowadays, the microcontroller board is inbuilt with a crystal oscillator as their source of the clock. The available repelling devices consist of a 12VDC battery-operated pulse generator, and a relevant speaker is used for the transmission [9]. Some electronic insect repelling system consists of the LCR oscillator, transistor-based amplifier, tripping circuit, sonic circuit, pre-amplifier, and frequency selection circuit modules [10]. However, these systems produce varying frequencies in most cases. Thus, continuous transmission of the stable frequency could be questionable. The consideration of the microcontroller-crystal oscillator combination counted for solving this problem. An AT89S52 microcontroller, ultrasonic generator, DC power supply, resonance matching detector, and power amplifier are the main functioning modules of such systems where IC IRFP460 is applied as a power amplifier [11]. Some microcontroller-based designs are configured with the Colpitts oscillator where stray reactance appears across the crystal degrades the performance. The performance is less affected in the Pierce configuration, where more accurate frequency, stability, and low power consumption crystal oscillators are used. Nevertheless, the accurate low-frequency crystal is relatively unavailable. Overdriving a crystal from unwanted high-frequency emission and power consumption may start at an overtone or fail to start at all, which are more common in available crystals [12]. With the combination of the programmable module, the simulated circuit system offers to drive a ($\geq$ 7.9 kHz) narrowband ultrasound signal. A trans-impedance pre-amplification was considered in this application for transducer excitation. Before using a power transformer, it needs to amplify the current rather than the voltage. The current signal has converted to the voltage signal for precision-level transformation. So, a trans-impedance amplifier (TIA) was introduced for this operation. The TIA bandwidth has been chosen wide enough. In this design, a resistive feedback trans-impedance amplifier was applied [13]. For the most transducer excitation techniques, transducer pre-amplifier was designed. Such kinds of pre-amps have a low signal-to-noise ratio [14]. The step-up transformer transforms the voltage by maintaining a 2:5 turn ratio. The transformed signals are expected to hit on the Capacitive Micro-Machined Transducer (CMUT) plate. This system recommends for driving the 82 kHz frequency-operated transducers [15].

The application of inaudible US is affected by the bio-environment. In the range of 20–40 kHz, the number of the cavitational spots caused by the high and low pressure of the wave becomes low, and energy associated with the bubble becomes higher. At higher frequencies (80–100 kHz), the number of spots becomes higher, and the energy releases become lower for the smaller bubble size. The frequency range of 80–100 kHz can be used on weevil, and it needed to know the frequency response before applying it. The *Sitophilus Oryzae* or rice weevil frequency response is not yet founded. A neurophysiological record says that the *beetle's* auditory system is sensitive to frequencies between 20 and 80 kHz [16]. The research work aimed to develop a health hazardless ultrasonic device. The output signal was analyzed from the simulation. This weevil repelling device is suggested for locating at the storage area by confirming the azimuth, altitude, and distance covered the area by radiation.

2 Method

2.1 Description of the Method

Ultrasonic oscillator design: For the sound operating speed, an Arduino-based parallel fundamental 30 PPM 20 pF load was chosen as a signal source. Across the board, a 5 V has been applied for powering up the circuit system. The inbuilt Arduino board is based on an ATMEGA16U2 microcontroller, 20 pF 50 V capacitor, 1 uF 100 V surface mount capacitor, 0.1 uF 100 V capacitor, 100 V 300 mA general purpose diode, 0.125 W 10,000 Ω metal thick film resistor, and 0.125 W 510 Ω surface mount resistor which makes the performance good enough. The circuit was designed on the circuit.io platform. Figure 1 describes the system circuit power connection. The GND, 5 V, V_{in}, and analog ports A0, A1 of the Arduino were applied in this system. The crystal resonator in Arduino Nano was programmed, which has a maximum clock speed of 16 MHz. A 32 KB inbuilt flash memory was used for the bootloader. The continuous frequency generation of Arduino Nano reduces the power. For boosting the signal, we had used the L293D H-bridge ready module. The output voltage of the module was 12 V. We installed the open-source Arduino IDE in lab PC to write the required code and upload programs to the board. Besides the software installation, we installed the relevant driver for operating the Nano. The Arduino sketch was performed most of its work in the setup() function of the program. In the code, all the ports of Nano were set to be as output. Timer1 had been configured for triggering an interrupt of 164 kHz. When each interrupts converted to the state of analog ports, it turned a 164 kHz signal into full-wave cycling at 82 kHz. The setup() function of the program has declared as: `OCR1A = 97;//Set compare register (16 MHz/97 = 164 kHz sine wave- > 82 kHz full-wave`. Directly, we had applied the L298N stepper driver board with Nano. The inbuilt L298N board has connected the load through a 100 nF capacitor and 1N4007 diode in parallel.

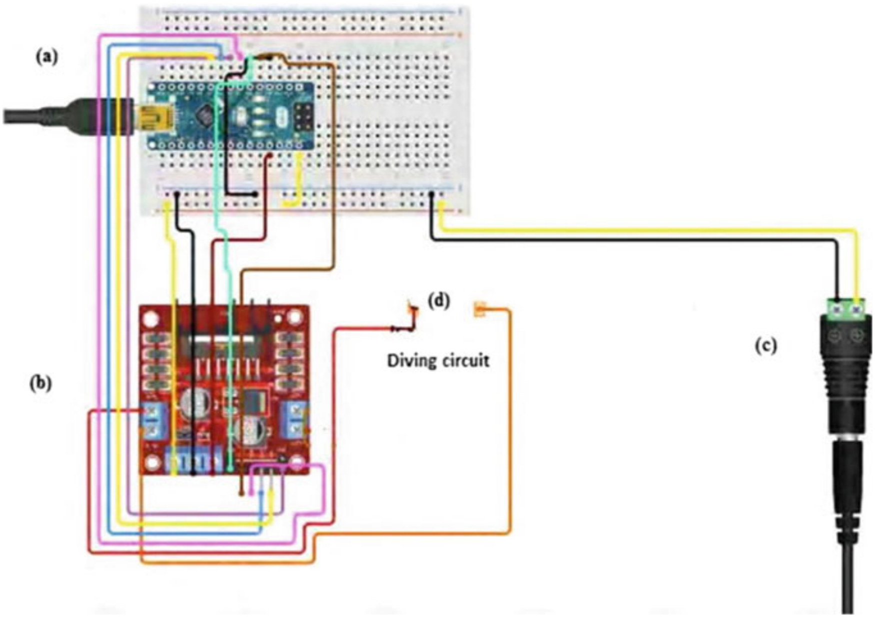


Fig. 1 Programmable oscillator applying with Arduino Nano, L298N 12 V driver module (the system has been designed and simulated on the circuito.io platform). **a** Represents the 16 MHz crystal oscillator in Arduino Nano powered from the L298N module. **b** L298N 12 V driver module, which has a 12 V output. **d** The Wein bridge oscillator is connected in series to the output of the driver module

The Wein bridge oscillator (RC-coupled) has oscillated with the pre-programmed frequency generation. The Wein bridge configuration was designed of an RC oscillator, op-amp, and feedback loop circuit. Generally, the bridge oscillator uses a feedback RC network of the same component values in parallel. Depending upon the frequency, the bridge configuration produces a phase delay or phase advance. At the resonant frequency, the phase shift is 0 without any externally applied input signal; the Wein bridge oscillator responses well at 1 kHz. The high stability configuration of the Wein bridge oscillator was shown a low distorted output at the resonant frequency. The RC oscillator combines with a bandpass filter connected to the positive terminal of the OPA 132. For a more precise oscillation, a resistive feedback element was connected to the inverting terminal of the op-amp. Such configuration turns about a 180-degree phase change. In Fig. 3, the resistors and capacitors values are applied as $C_1 = C_2$ and R_1 almost equal to R_{36}. When the oscillator signal amplitude increases, it effectively decreases the value at the output, and the resistance decreases. For the limiting amplitude of the OPA132, the BZV80 diode was applied parallel to the feedback resistor. The diode was used at inverting terminal of the op-amp, which was placed across the pair of resistors. OPA132 had been chosen for a maximum of 50 pA input which shows a wide BW of 8 MHz, low distortion: 0.00008%, and a wide supply range of ±2.5 to ±18 V. This oscillator circuit limited the output to 6.51 V. During the signal propagation through the circuit components, it might have increased temperature. The BZV80 shows a low-temperature coefficient range which has a maximum of 0.01%/K. The output frequency was calculated by using the formula [17]:

Frequency of the oscillation:

$$f = \frac{1}{2\pi\sqrt{R_1 R_{36} C_1 C_2}}$$

Let us assume, $R_1 = R_{36}$ and $C_1 = C_2$ $C_1 = C_2$

$$f = \frac{1}{2\pi R_1 C}$$

A 10 nF capacitor and a resistor of 1 kΩ were chosen to match the resonance frequency of 82 kHz because of the unavailable 0.194 kΩ resistors in the following simulated circuit.

Voltage source VG1 in Fig. 2 was applied here only for simulating purposes. The IN4148 diode was applied at the oscillator's output, with a maximum current carrying capacity of 300 mA. The diode can peak up to 2 A current and has the specialty of a fast recovery time of 8 ns at a forward current of 10 mA. This might have shown dynamic characteristics at the initial point, but it has shown static characteristics or DC resistance during specific current flow time.

Signal amplification and narrow bandpass circuit: The signal shape can be lost during the signal transit through the oscillator. For this reason, we were chosen a 15 MHz BW LM318 op-amp. A supply voltage range operates this op-amp from

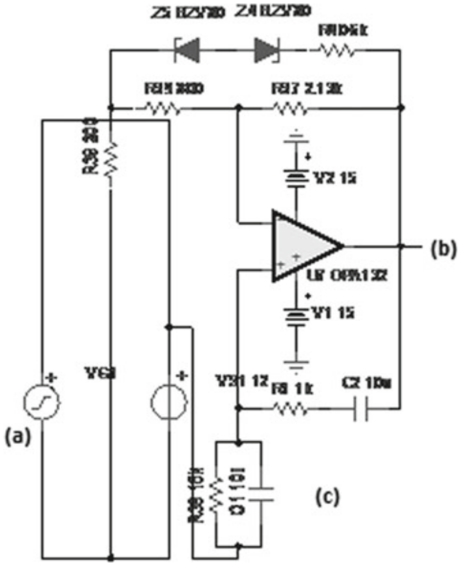


Fig. 2 Wien bridge oscillator for 82 kHz frequency designed in TINA-TI platform. **a** The voltage source has been replaced with the previous output of the L298N driver module. **b** The Wein bridge oscillator output has been collected, which was designed to consider a feedback loop to the input. **c** Represents a tank circuit that oscillates with the matching of the resonance of the programmatic oscillator

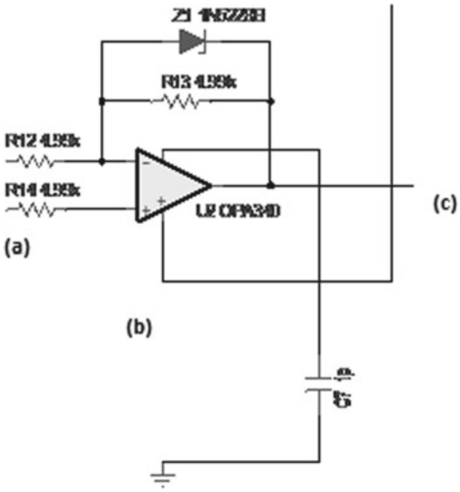


Fig. 3 Trans-impedance amplification has been applied for the current to voltage conversion of the processing signal to drive an ultrasonic transducer. **a** The output of the symmetrical circuit acts as an input of the trans-impedance circuit. **b** Applying a high-speed rail-to-rail CMOS operational amplifier with an excellent dynamic response (BW = 5.5 MHz and quiescent current is only 750 μA). **c** Trans-impedance output

±5 V to ±20 V. We have used LM 318 op-amp for internal frequency compensation features for our 82 kHz frequency drive. This op-amp-based circuit had been designed as inverting input according to the input resistor and feedback resistor. LM 318 op-amp was used for amplification of the output voltage by a factor of gain ratio. For overcoming the variation of frequency from the oscillator, a narrow bandpass filter has been introduced. The output of this circuit has a center frequency of 82 kHz. We have applied both positive and negative feedback loops. Such configuration has the advantage of any change in the most significant gain at the center frequency.

Maintaining the filter circuit quality factor as Q > 10 is called a narrow bandpass filter [18]. When BW increases, the quality factor decreases, and when BW decreases, the quality factor increases. The quality factor as:

$$Q = \frac{f_c}{BW}$$

Full-wave rectification is the method of rectification used for every half cycle of the input signal instead of the other half cycles. For both input cycles, a unidirectional current-based circuit configuration was designed by applying 1N4148 diodes. Considering the low cost and good signal output, this full-wave rectifier was shown a uniform output signal. The output has a unidirectional signal which is doubled of the half-wave rectifier. With considering no signal loss, the average DC output can be calculated from the equation derived below [19]:

$$V_{\text{dc}} = \frac{2V_{\max}}{\pi} = 0.637V_{\max}$$

At the output of the rectifier circuit, a capacitor was applied for smooth DC output voltage.

Standardize symmetrical output: 4013 IC was configured for the rectifying circuit output. This IC consists of two identical, independent data type flip-flop. One flip-flop consists of the pins like data, set, reset, clock inputs, and outputs. We were connected $2\overline{\text{Q}}$ to the data input 2D of the second flip-flop of the same IC for counter application. During the positive-going transition of the clock pulse, 2D input was transferred to the $2\overline{\text{Q}}$ output. This operation had set and reset the device and was accomplished when the signal level was high. The 4013 considers an input current of 1 μA at 18 V and has a noise margin of 1 V at $V_{DD} = 5$ V, 2 V at $V_{DD} = 10$ V, 2.5 V at $V_{DD} = 15$ V. The symmetrical circuit can drive by the higher range of low frequencies when occurred [20]. When the rectifier module achieves the double of the center frequency, the symmetrical module divides the frequency by 2.

Trans-impedance amplification: The OPA340 op-amp was selected for current to voltage amplification purposes in this part of the circuit system. This op-amp can be optimized for low voltage, single-supply operation. As an ideal transformer, operation was considered in the design (Fig. 3) so that I/V converted analog signal as an input of the transformer is highly needed for smoothing the operation.

The symmetrical signal was operated through the OPA340 instrumentation. The trans-impedance circuit was designed as a buffer for the symmetrical signal and acted as digital to analog converters (DAC). According to Ohm's law, the feedback resistor and input current can be derived as [21]:

$$V_{\text{o}} - V_{\text{inv}} = R_{13} I_{\text{i}}$$

According to the op-amp law,

$$V_o = a(0 - V_{inv}) = -aV_{inv}$$

where a is the open loop gain,

$$A = \frac{V_o}{I_i} = \frac{R_{13}}{1 + \frac{1}{a}}$$

Here, A is the trans-impedance gain.

Ultrasound pre-amps instrumentation: For ultrasound transmission, we had chosen the LMH6629 op-amp. The pre-amplification circuit can operate both DC and AC signals. This instrumentation was built with this ultra-low noise, a high-speed operational amplifier [22]. This op-amp has low-input noise ranging from 0.69 $\frac{\text{nV}}{\sqrt{\text{Hz}}}$ to 2.6 $\frac{\text{pA}}{\sqrt{\text{Hz}}}$ and low distortion of $-90 \sim 94$ dB, and ultra-low DC errors. The decision was to choose this I.C. for precise level amplification to drive the transformer by maintaining the same input range. The LMH6629 has excellent offset voltage. Due to the input bias current, it avoids voltage errors. In this non-inverting circuit configuration (Fig. 4), a parallel combination of a gain setting resistor R_{17} and feedback resistor R_3 value was determined. R_3 was selected as an almost equal value of the source resistor of the non-inverting terminal's resistor of R_{s16}. The following equation can determine the value of the resistors of R_3 and R_{17}:

$$R_3 = A_v R_{s\,16}$$

$$R_{17} = R_3/(A_v - 1)$$

$$A_v = 1 + \frac{R_3}{R_{17}}$$

Power transformer for transducer excitation: An ideal step-up transformer (Fig. 5) was selected, maintaining a 2:5 turn ratio of the primary and secondary winding [23]. The ratio can be expressed as:

$$N_p : N_s = 2 : 5$$

Here, N_p is the number of turns in the primary coil, and N_s is the number of turns in the secondary coil, primary voltage is V_{F3}, the primary current is I_{F3}, the secondary voltage is V_{F4}, and secondary current is I_o.We calculated the input–output current and voltage by using the equation as follows [24]:

$$\frac{N_p}{N_s} = \frac{I_o}{I_{F3}} = \frac{V_{F3}}{V_{F4}}$$

$$I_o = \frac{N_p}{N_s} I_{F3}$$

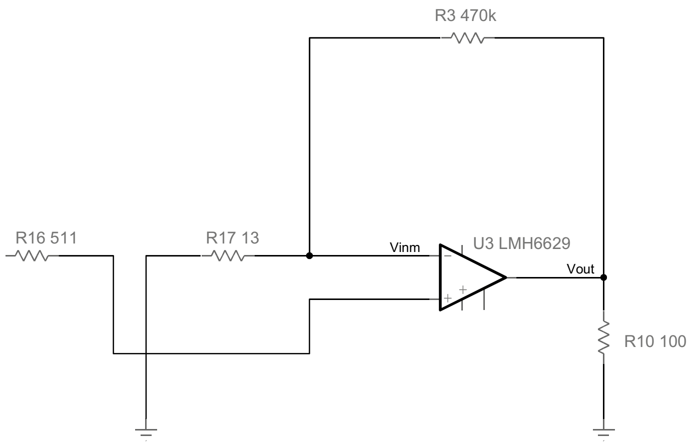
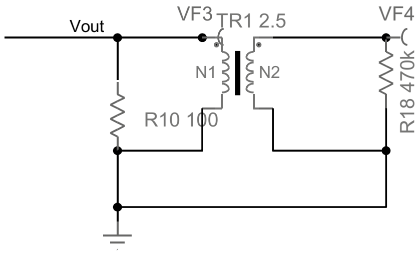


Fig. 4 Low-noise pre-amplifier for transducer drive

Fig. 5 Transformer circuit for transducer drive



A specific transducer can operate appropriately due to the maximum excitation. We were given priority for matching the frequency of the driving circuit and the transducer. The frequency generator might be pretty unstable. A range of varying frequencies might be generated at the hardware level. Because of that, a variable ranged frequency driving transducer has been selected, which can drive an entire frequency range of 30 kHz to 300 kHz. The transducer's output voltage specifies 200–500 V_{pp} and is adjustable with the defined range of frequencies [25]. The supply voltage range of the transducer is 12–24 VDC.

3 Result and Discussion

The microcontroller-based oscillator acts as the pure sin wave function generator. The simulation circuit was designed for Wien bridge configuration only, making the pre-determined function generator's frequency resolution (82 kHz). We had followed the Changpuakonline calculator to determine the output frequency of the Wien bridge circuit for different resistors and capacitors values. For a clear 82 kHz frequency

Table 1 Calculating different values for Wien bridge frequency response

Resistor values (kΩ)	Capacitor values (nf)	Frequencies (kHz)
1.326	2	60
520.965	0.0047	65
483.754	0.0047	70
2.2	1	72
42.44	0.050	75
40.809	0.050	78
1.989	1	80
1.2	1.5	82
0.194	10	82
0.959	2	83
1.872	1	85
1.768	1	90

resolution, we had determined the resistor value as 1.2 kΩ and capacitor value as 1.5 nf. Table 1 has given different R_1 (kΩ) and C_2 (nf) values for different frequencies. The desired frequency can be changed according to the matching of the function generator. The simulated circuit can be responded to in the range of 60–90 kHz frequency when the different R_1 (kΩ) and C_2 (nf) values are changed with the desired manner in the Wien bridge circuit configuration. The selected frequency (kHz) results for variable resistors and capacitors have been given in Table 1 [26]. For fulfilling the need, we had chosen the resistors and capacitors in the Wien bridge configuration, and it is highly recommended that the frequency should be ranged from 60 to 90 kHz.

3.1 Different Low-Frequency Output from the Same Simulated Circuit

The output from the oscillator circuit was recognized as VF5, which shows an 82 kHz frequency AC output. The arbitrary values of the Wien bridge resistor and capacitor had taken from above Table 1. For the values of $R_1 = 2$ kΩ and $C_2 = 1$ nF or $R_1 = 1.2$ kΩ and $C_2 = 1.5$ nF or $R_1 = 0.194$ kΩ and $C_2 = 10$ nF, the following results were found from the simulation. The simulation was turned the desired result by changes of the values of the oscillator circuit only. The RC oscillator circuit was also simulated for 90 kHz frequency but had a good frequency response in the resonance condition of 82 kHz (Fig. 6).

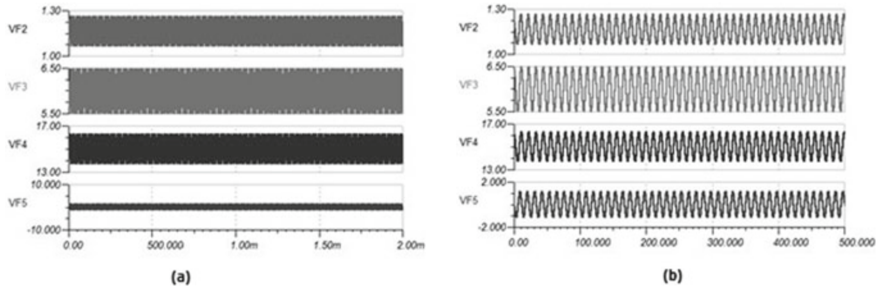


Fig. 6 Oscillator output signals (left to right: 82 and 90 kHz). **a** In the picture, the 82 kHz frequency simulated output had achieved for the defined resistors and capacitors values. **b** Represented a 90 kHz output signal driving according to the same RC values

3.2 DC Steady-State Analysis of the Oscillator and Final Output Signal:

The steady-state analysis of zero initial value and the transient method was considered for the simulation. The results were found with 90% accuracy, 1% absolute value increment, and maximum 100 mA current increment. The capacitor threshold 100 pF and iteration limit of 1000 were defined during the simulation. The VF_4 in Fig. 7 represents a clear 15 V DC signal where VF_5 is expressed as the oscillator signal.

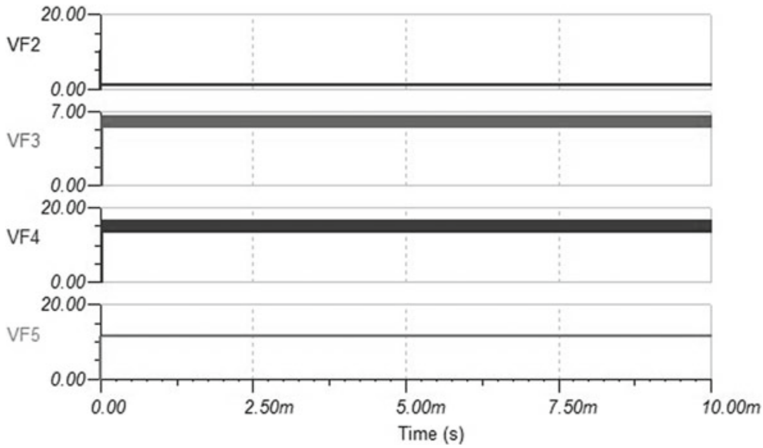


Fig. 7 DC steady-state signal output of the oscillator (VF_4) compared with final output (VF_5)

3.3 Nodal Analysis and Trans-Impedance Gain Calculation:

The oscillator output voltage was found as 11.47 V. It is needed to reduce the voltage for the proper operation of the narrow bandpass circuit. So, we had applied a rectifier and a resistor for limiting voltage input at the narrow bandpass filter as 5.99 V. The full-wave rectifications were applied before the operation of the D-type flip-flop (4013). Through the rectification, it had a voltage drop from 1.61 V to 30.51 mVA parallel capacitor has been used for smoothing the rectification output. The D-type of 4013 IC had connected to the circuit, which generated a 50 mV balanced output. The differential input was applied consecutively to the negative terminal and at the positive terminal of the OPA 340-based trans-impedance circuit as 2.61 and 5.3 V. The positive terminal input current was measured as 425.13 μA, which had a simulated output of 3.22 V.

Here,

$$V_o = 3.22\,\mathrm{V},\ V_{\mathrm{inv}} = 2.61\,\mathrm{V},\ R_{13} = 4.99\,\mathrm{k}\Omega,\ I_i = 425.13\,\mu\mathrm{A}$$

According to the op-amp law, the derivation is-

$$V_o = a(0 - V_{\mathrm{inv}}) = -aV_{\mathrm{inv}}$$

$$a = \text{open loop gain} = -\frac{V_o}{V_{\mathrm{inv}}} = -\frac{3.22}{2.6} = 1.238$$

A is the trans-impedance gain:

$$A = \frac{V_o}{I_i} = \frac{R_{13}}{1 + \frac{1}{a}} = \frac{4.99}{1 + \frac{1}{1.238}} = \frac{4.99}{1.807} = 2.76$$

For the high gain amplification ($A_v = 10$) and low-noise (−3 dB) operation, LMH6629 op-amp-based pre-amp instrumentation has been considered to drive the power transformer.

3.4 Transformer Output Analysis

The simulated output for the ultrasound pre-amp instrumentation circuit was identified as the primary voltage and current, which was simulated as $V_{F3} = 6\,\mathrm{V}$, $I_{F3} = 60$ mA. From the transformer equation, the output of the overall circuit system had been calculated as:

$$\frac{N_p}{N_s} = \frac{I_o}{I_{F3}} = \frac{V_{F3}}{V_{F4}}$$

$$I_o = \frac{N_p}{N_s} I_{F3} = \frac{2}{5} * 60\,\text{mA} = 24\,\text{mA}$$

$$\frac{N_p}{N_s} = \frac{V_{F3}}{V_{F4}} \text{ or}$$

$$V_{F4} = \frac{N_s}{N_p} V_{F3} = \frac{5}{2} * 6 = 15\,\text{V}$$

3.5 DC Transfer Characteristics

The input signal of the circuit system was applied as a pure sin wave which is an AC signal. We had designed a D-type flip-flop circuit that contained a low signal rather than the input for further signal operation. From Fig. 8, the DC transfer characteristics have shown the input voltage Vs output. The upper most signal was represented as the transformer output. Then, the immediate lower signal has shown here, expressed as VF5 or oscillator output. The immediate lower signal was identified as VF3, which has shown transformer input signal or trans-impedance instrumental circuit gain output.

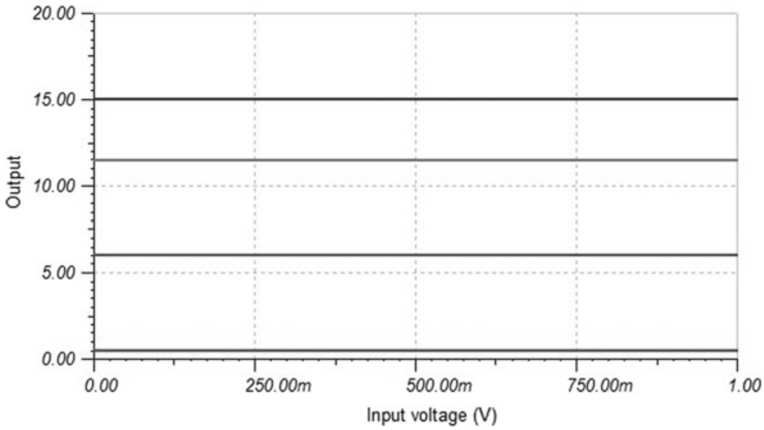


Fig. 8 DC transfer characteristics of trans-impedance output

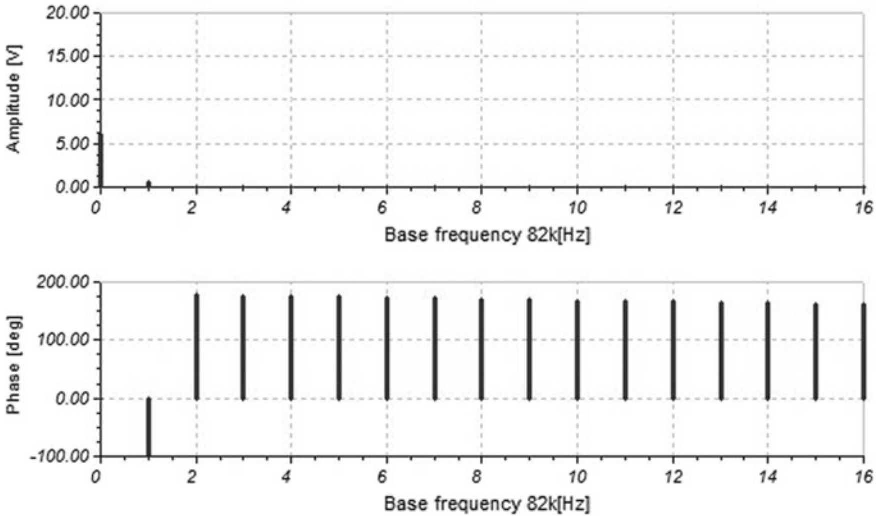


Fig. 9 Fourier coefficients for 82 kHz frequency signal in the frequency domain (frequency vs. phase and amplitude)

3.6 Fourier Analysis and Harmonic Distortion of Overall Circuit System:

We have simulated all operating points for Fourier series analysis. The analysis represents the signal in the time domain for 82 kHz frequency. Fourier coefficient analysis (Fig. 9) in the simulation process was considered for 16 harmonics and 4096 samples for the signal expressed as $D * \cos(k\omega t + \theta)$. Table 2 represents the time, amplitude, and phase change of the final output. From the simulation results of the circuit system, 0–16 harmonics have shown a phase change of $-18.54°$ of the originated signal from the oscillator.

Harmonic distortion at VF_4 (transformer's output) had simulated as 0.2463%.
Harmonic distortion at VF_5(oscillator output) had simulated as 306.15%.
Harmonic distortion at VF_3 (transformer's input) had simulated as 0.2463%.
Harmonic distortion at VF_2(D-type flip-flop) had simulated as 343.98%.

3.7 Result from Transient and Hysteresis Analysis of DC Outputs

The circuit simulation responses had measured from 0 to 100 μs (Fig. 10). For every time step, all the nodal voltage and current were measured from the simulation. These waveforms have been defined using pair of time–voltage coordinates.

Table 2 Fourier series coefficients

Time	Amplitude	Phase
0	15	0
1	1.25	−89.96
2	811.35μ	−2.64
3	809.46μ	−3.93
4	807.05μ	−5.18
5	804.28μ	−6.38
6	801.3μ	−7.55
7	798.34μ	−8.68
8	795.52μ	−9.77
9	792.97μ	−10.85
10	790.74μ	−11.91
11	788.81μ	−12.97
12	787.12μ	−14.04
13	785.58μ	−15.13
14	784.04μ	−16.24
15	782.38μ	−17.38
16	780.48μ	−18.54

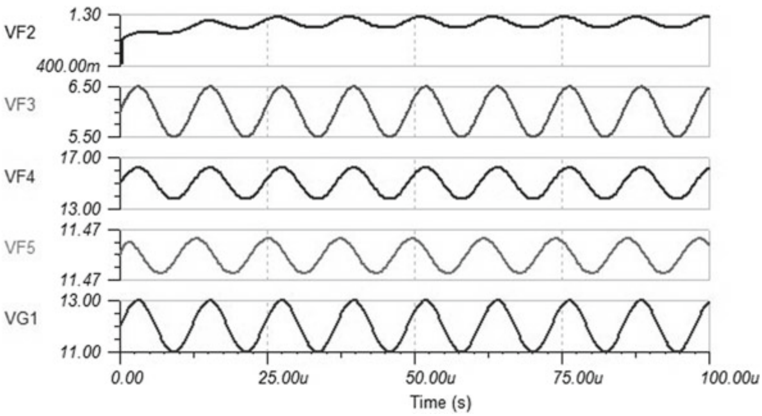


Fig. 10 Mixed transient analysis (final output VF4)

The output signal's hysteresis analysis shows in Fig. 11, which had started from 1 to 15 V of 100 points. The lower signal from the D-type flip-flop was recognized as VF2. The transformer input signal was recognized as VF3, which has increased due to the increase of input voltage in a linear way. When the transformation has occurred, the output signal increases as long as the increase of input voltage. The transformer's output was shown as VF4, which have recognized as the upper signal

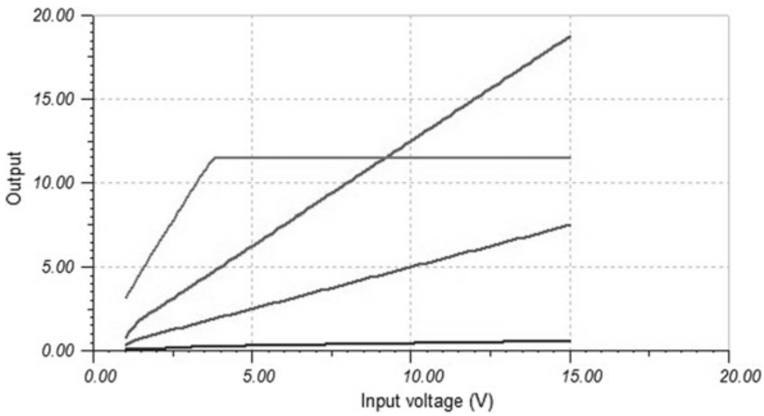


Fig. 11 Hysteresis analysis of DC outputs (input voltage vs. output voltage)

of VF3. We had maintained a 2:5 turn ratio for the transformer in the simulation process. For 12 V input, it was reached the maximum of 15 V. The oscillator output voltage had maintained as approximately 12 V.

The system described here is a design and simulation only. Here, an ideal operation was maintained in the simulation process. Due to the propagation of the signal, the temperature has increased nonlinearly.

3.8 Discussion of the Results

We had considered a full sin wave with a 0-degree phase angle for 12 V input voltage in the simulation process. The system could be deployed by a 12 V battery on the replacement of the voltage source. The oscillator circuit had dual input at inverting and non-inverting terminals. The inputs of the circuit system were measured as 3.76 V and 12 V. Both the op-amp terminals have a feedback connection with their output. The non-inverting terminal was created oscillation for the RC values that already have discussed. The 5.99 V output voltage from the oscillator was driven by rectification to the input at the LM318 inverting terminal. A precise level output was shown as 6 V. A narrow bandpass filter was applied to integrate with the LM318 inverting terminal. The input of 1.6 V consisted of a series of 100 Ω resistor and 1 μF capacitor and feedback of 1 kΩ resistor to the output. A full-wave rectifier has doubled the frequency in its DC output and divides it through a D-type flip-flop. The symmetrical circuit input was measured as 30.51 mV. The outputs from 2 and $2\overline{Q}$ were measured consecutively as 50 mV, 426.61 μA, and 50 mV, 425.13 μA. These output signals have symmetrical characteristics. The trans-impedance amplification was applied at the output of the D-type flip-flop. OPA340 CMOS op-amp has been applied in this circuit system with dual input of 2.6 V at inverting terminal and 5.3 V at the non-inverting terminal. The output of pre-amp instrumentation has given a high-precision output of 3.22 V signal. The 5.3 V, 407.48 mA input at OPA340-based trans-impedance circuit output was applied at inverting terminal of LMH6629. The output of LMH6629-based pre-amp circuit was measured with a clear 6 V, 60 mA output. The 6 V was applied to the 2:5 turns ratio-based ideal transformer. The power transformer had transformed the signal to 15 V. The ultrasound transducer has to operate in the range of 12–24 V and transmits a voltage of 400 V_pp.

4 Conclusion

Affecting grain quality and grain losses by insects/pests in developing countries recognizes as a significant factor. With the reduction of grain losses, a novel weevil repeller circuit was designed and simulated. A semi-circular transmitter was considered for the transmission. The antenna has been conceded of the most relevant ultrasonic sensors. We have anticipated that the transmission of an 82 kHz frequency of US can prevent an undesired attack by the weevil on grains and seeds. The quality factor of the filter circuit was maintained. A trans-impedance gain of 2.76 has been maintained by choosing low noise and precision-level ultrasonic driving IC. Finally, the expected outcome was measured by simulation of a 60 mA signal. The Fourier coefficients were represented the clear phase change. The harmonic distortion of the oscillator was improved (306.15–0.2463%) in such kind of circuit configuration. An ideal transformer transposed the output as 15 V when the nodal analysis was executed during the simulation. The silent features of our electronic pest repel device ensure available power, the affordable cost for individual farmers, compactness, and no harm and toxicity, even considering the children's safety. The aim of further research will be an evaluation of the capabilities of the US.

References

1. Kisimoto, R., Dyck, V.A.: Climate and rice insects. In: Banos L (ed.) Proceedings of the Symposium on climate and Rice; International Rice Research Institute, Philippines, Climate and Rice. Los Ba±os, pp. 367–390 (1976)
2. Ahmedi, F., Mcloughlin, I., Chauhan, S., Haar, T.G.: Bio-effects and safety of low intensity, low-frequency ultrasonic exposure. Progr. Biophys. Mol. Biol. **108**(3), 119–138 (2012). https://doi.org/10.1016/j.pbiomolbio.2012.01.004
3. FAO: Insect Damage: Post-Harvest Operations (2017). Available online http://www.fao.org/3/a-av013e.pdf. Accessed 12 Dec 2017
4. Murdock, L.L., Seck, D., Ntoukam, G., Kitch, L., Shade, R.E.: Preservation of cowpea grain in sub-Saharan Africa—Bean/Cowpea CRSP contributions. Field Crop Res. **82**, 169–178 (2003)
5. Liu, X., Cui, X., Ma, F.: Design of an intellectualized ultrasonic generator. In: International Conference on Education Technology and Social Science, ICETSS 2014 (2014)
6. Awal, M.A., Ali, M.R.., Hossain, M.A., Alam, M.M., Kalita, P.K., Harvey, J.: Hermetic bag an effective and economic rice storage technology in Bangladesh. Paper presented at 2017 ASABE

Annual International Meeting, Spokane, United States (2017). https://doi.org/10.13031/aim.201700329
7. Kansu, Y.: A Transimpedance Amplifier for Capacitive Micromachined Ultrasonic Transducers (2015)
8. Galandanci, G.S.M., Ewansiha, K.O.: Design and Simulation of a 20 kHz to 50 kHz Variable Frequency Oscillator (VFO). IJRRAS **16**(1) (2013)
9. Ibrahim, A.G., Oyedum, O.D., Awojoyogbe, O.B., Okeke, S.S.N.: Electronic pest control devices: a review of their necessity, controversies and a submission of design considerations. Int. J. Eng. Sci. (IJES) **2**, 2319-1805
10. Singh, R.J.S., Hiremani, T.: Electronic Pest Repellent. EE318 Electronic Design Lab, Project Report, EE Dep t, IIT Bombay, April (2009)
11. Yturralde, K.M., Hofstetter, R.W.: Efficacy of commercially available ultrasonic pest repellent devices to affect behavior of Bed Bugs. J. Econ. Entomol. (2012)
12. Schreck, C.E., Webb, J.C., Burden, G.S.: Ultrasonic devices: evaluation of repellency to cockroaches and mosquitoes and measurement of sound output. J. Environ. Sci. Health Part A—Environ. Sci. Eng. **19**, 521–532 (1984)
13. Svilainis, L., Motiejûnas, G.: Power amplifier for ultrasonic transducer excitation. Ultragarsas **58**(1). ISSN 1392-2114 (2006)
14. Svilainis, L., Puodiunas, V.: Ultrasonic NDE system: the hardware concept. Ultragarsas **29**(1). ISSN 1392-2114, pp 34–40 (1998)
15. Cheeke, J.D.N.: Fundamentals and Applications of Ultrasonic Waves. https://books.google.com/books?id=Lnzb4PeTbeYC&pgis=1. Accessed 10 Feb 2018
16. Forrest, T.G., Read, M.P., Farris, H.E., Hoy, R.R.: A tympanal hearing organ in scarab beetles. J. Exp. Biol. **200**, 601–606 (1997)
17. Electronics-Tutorials. https://www.electronics-tutorials.ws/oscillator/wien_bridge.html
18. TI. https://www.ti.com/tool/TINA-TI
19. Mao, W., Zhang, Z.G.W., Zhu, R.J.X.: Design of a narrowband adaptable filter in low-frequency domain. J. Phys.: Conf. Ser. **1621**, 012085 (2020)
20. Basa, K.C., Gomez, K.S., Navarro-Tantoco, F.B., Quinio, A.S., Arada, G.P., Co, C.B.: Design of a varying ultrasonic frequency amplifier. In: IEEE Region 10 Annual International Conference, Proceedings/TENCON (2012). https://doi.org/10.1109/TENCON.2012.6412169
21. TI. https://www.ti.com/lit/ds/symlink/cd4013b.pdf
22. TI. https://www.ti.com/lit/an/sboa268a/sboa268a.pdf?ts=1612682731905%26ref_url
23. Dumbrava, V., Svilainis, L.: Noise model for ultrasonics transducer preamplifier. ElektronikaIrElektrotechnika **66**(2), 26–30 (2006). Retrieved from https://eejournal.ktu.lt/index.php/elt/article/view/10587
24. Toppr. https://www.toppr.com/ask/en-bd/question/in-a-stepup-transformer-the-ratio-of-turns-of-primary-to-secondary-coils-is-110/
25. Lkouniv. https://www.lkouniv.ac.in/site/writereaddata/siteContent/202004101310174191gaurav_gupta_Transformers.pdf
26. Koehler, P.G., Patterson, R.S., Webb, J.C.: Efficacy of ultrasound for German Cockroach (*BlattellaGermanica*) (Ortoptera:Blattellidae) and Oriental Rat Flea (*Xenopsyllacheopis*) (*Siphonaptera: Pulicidae*) control. J. Econ. Entomol. **79**, 1027–1031 (1986)

Exploration of Electronic Structure and Optical Properties Hafnium (IV) Oxide and Its 12% Si, Ge and Sn with Material Data Analysis by Computational Methods

Unesco Chakma, Ajoy Kumer, Tomal Hossain, Md. Sayed Hossain, Md. Monsur Alam, Nusrat Jahan Khandakar, Md. Shariful Islam, Rubel Shaikh, and Md. Hazrat Ali

Abstract This research work conveys the computational analysis of hafnium (IV) oxide and its doped crystal by Si, Ge and Sn replacing on the oxygen atom in HfO_2, as hafnium (IV) oxide has been used in power-electronics devices of MOSFETs and electronics as RRAM due to wide band gap which makes a vast problems creating high resistances. Regarding this case, the hafnium (IV) oxide has selected inputs

U. Chakma · T. Hossain · Md. Hazrat Ali
Department of Electrical and Electronics Engineering, European University of Bangladesh, Gabtoli, Dhaka 1216, Bangladesh
e-mail: unescochakma@gmail.com

Md. Hazrat Ali
e-mail: dr.hazrat_ali@eub.edu.bd

U. Chakma · A. Kumer (✉) · Md. Monsur Alam
Laboratory of Computational Research for Drug Design and Material Science, Department of Chemistry, European University of Bangladesh, Gabtoli, Dhaka 1216, Bangladesh
e-mail: kumarajoy.cu@gmail.com

Md. Monsur Alam
e-mail: munna@eub.edu.bd

A. Kumer
Department of Chemistry, Bangladesh University of Engineering Technology, Dhaka 1000, Bangladesh

Md. Sayed Hossain
Center for Research Reactor, Bangladesh Atomic Energy Commission, Dhaka, Bangladesh
e-mail: sayedrj34@gmail.com

Md. Monsur Alam · N. J. Khandakar · Md. Shariful Islam
Department of Civil Engineering, European University of Bangladesh, Gabtoli, Dhaka 1216, Bangladesh
e-mail: eubsharif@gmail.com

R. Shaikh
Department of Computer Science and Engineering (Division of ICT), European University of Bangladesh, Gabtoli, Dhaka 1216, Bangladesh
e-mail: rubel@eub.edu.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_18

how the band gap has decreased later than doping by the large surface area atoms, such as Si, Ge and Sn. The first principle method in view of density functional theory (DFT) expresses the structural geometry, electronic structure and optical properties utilizing conformist calculations pertaining to HfO_2 executing the computational avenue of the CASTAP regulation from material studio 8.0. The band gap was found in 4.340 eV, 2.033 eV, 1.686 eV and 3.210 eV for HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals through the Generalized Gradient Approximation (GGA) with Perdew Burke Ernzerhof (PBE), and the DFT and PDOS were simulated for evaluating the nature of $6s^2$, $5p^6$, $4f^{14}$, $5d^2$ orbital for a Hf atom, $3s^2$, $2p^6$ orbital for Si atom, $4s^2$,$3p^6$, $3d^{10}$ orbital for Ge atom, $4d^{10}$, $5s^2$, $5p^2$ for Sn atom, 2s and 2p orbital for O atom of $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals. The optical properties, for instance, absorption, reflection, refractive index, conductivity, dielectric function and loss function, were calculated.

Keywords Band gap · DOS · Optical properties · Dielectric function · Doping · DFT

1 Introduction

On the scientific area, there are many challenges for the use of the existing and prospect generations of MOSFETs [1, 2]. MOSFETs consist of four parts, such as body, gate, drain and sources, while a decisive and instantaneous barrier is occurred during the optimization of the gate stack. In this case, the interfaces of HfO_2 with Si or metal electrodes can be enhanced by optimization and operation of MOSFETs in opto-electric devices, while the Fermi level of the metal is remained between minimum conduction band (MCB) and maximum valence band (MVB) of Si for the application in RRAM and ferromagnetic devices [2, 3]. Next, the higher viable opportunities for MOSFETs are involved for Si and non-Si substrates (Ge and Sn) with HfO_2 in the thin film and edge CMOS applications [4]. It has been illustrated that the additional and elementary modeling exertions of molecular crystals for HfO_2, and doped by Si, Ge and Sn through the analysis of first principle methods using computational theory has executed in this study that is almost zero before. The main point of doping replacing the oxygen atom in HfO_2 is figured out for the valance shell and its structural configuration although the main driving force comes from the metal atom for MOSFETs. In addition, it has accounted for the other impending applications in resistive random access memory (RRAM) [5], automotive electronics, power converter, insulated gate in the bipolar transistor and amplifier of electronic devices [6–8]. Due to poor conductivity of oxygen atom in HfO_2, a large defect and empirical problem have been raised for the stored information in filaments, nano-ionic devices and grain boundaries [9, 10] even opto-electronic devices [11]. For the case, it is the strongest and foremost fact for investigation of replacing the oxygen atom in HfO_2 by doping to produce new doped crystal as new MOSFETs [12] for uses in RRAM [13] and other electronic devices. On the other hand, it is mentioned and substantiated that the doping is a widely utilized method to overcome the barrier from MVB to MCV of electron transferring for recovering and proliferating the performance of metal oxide materials. However, this manuscript has sorted out as follows HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals including the electronic structure, density of states and optical properties. For evaluating the required parameters, the GGA with PBE avenue of computational method is used due to its outstanding accuracy and acceptance for heavy metals containing crystals [14–22]. Finally, a comparative study for reflectivity, absorption and dielectric function among HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ has performed for further investigations.

Lastly, in the perspective of our country, Bangladesh is going to face the 4th Industrial Revolution in both the industries and advance electronics material sectors, where adequate MOSFETs, resistive random access memory (RRAM) characteristics of HfO_2-based MOSFETs and RRAM device resources and development will be necessitated, otherwise, the revolutions will be hindered due to energy and lost the achievement for the goal of next development. To solve the upcoming difficulties with advanced electronics materials, the HfO_2 is one of the most demanded materials. Unfortunately, the wide band gap (5.69 eV) makes many difficulties for their versatile uses in opto-electronic devices. Regarding this case, the silicon (Si), germanium (Ge) and tin (Sn) are doped to form as $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals which are compared with HfO_2 and take the challenge for the 4th Industrial Revolution in our country.

2 Computational Methods

For case of simulation and optimization, there are placed some conditions, such as the convergence criterion at 3×10^{-6} eV/Å, total energy at 1×10^{-3} Å and the maximal stress at 5×10^{-2} GPa. This condition has executed for geometry optimization by GGA with PBE functional which is taken from CASTEP module of the material studio V8.0 [23]. In addition, the cut-off for the molecular optimization and structural geometry was obtained at 500 eV, and k point was remained at $2 \times 2 \times 2$ having the setting of norm-conserving pseudo-potentials options for evaluation of the electronic structure, band structure and structural geometry. Next, the optical features were similarly simulated by the same condition and function, as well as a comparative data for the band gaps, DOS, PDOS and optical properties for HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$.

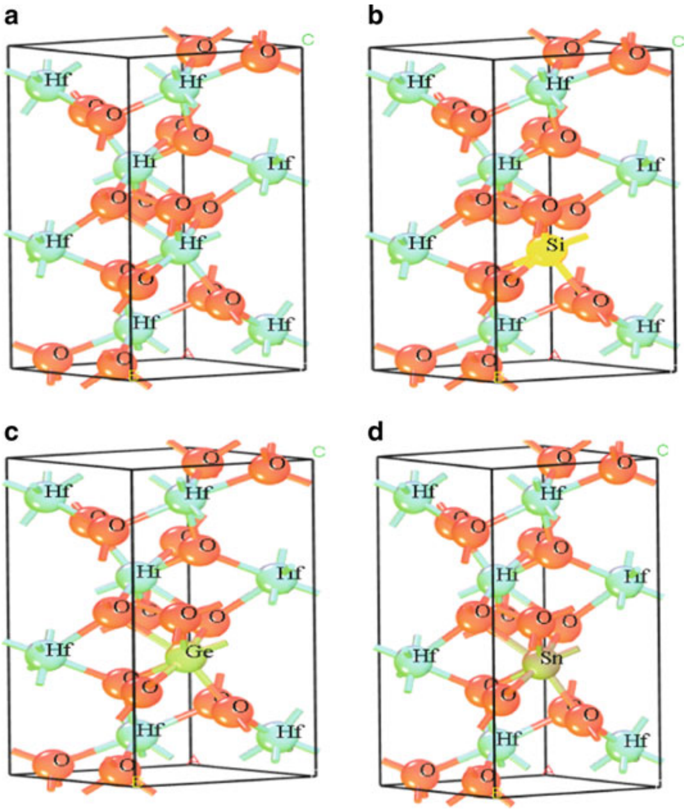


Fig. 1 **a** Structure for HfO_2, **b** structure for $Hf_{0.88}Si_{0.12}O_2$, **c** structure for $Hf_{0.88}Ge_{0.12}O_2$, **d** structure for $Hf_{0.88}Sn_{0.12}O_2$

Table 1 Structural calculation by GGA with PBE

Crystals	a	b	c	α	β	γ	Crystal type	Space group	Density g/cm^3
HfO_2	5.040	5.074	5.269	90.00	90.00	90.00	Orthorhombic	$Pca2_1$	10.37
$Hf_{0.88}Si_{0.12}O_2$	5.040	5.074	5.269	90.00	90.00	90.00	Orthorhombic	$Pca2_1$	10.37
$Hf_{0.88}Ge_{0.12}O_2$	5.040	5.074	5.269	90.00	90.00	90.00	Orthorhombic	$Pca2_1$	10.37
$Hf_{0.88}Sn_{0.12}O_2$	5.040	5.074	5.269	90.00	90.00	90.00	Orthorhombic	$Pca2_1$	10.37

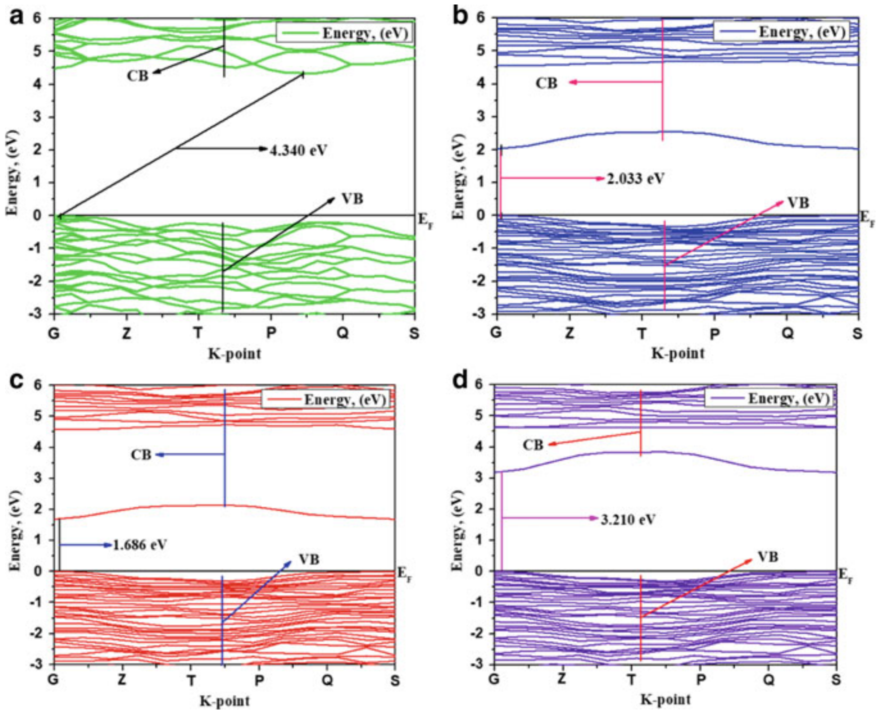


Fig. 2 **a** Electronic structure HfO_2, **b** electronic structure $Hf_{0.88}Si_{0.12}O_2$, **c** electronic structure $Hf_{0.88}Ge_{0.12}O_2$, **d** electronic structure $Hf_{0.88}Sn_{0.12}O_2$

3 Results and Discussions

3.1 *Geometry of Optimized Structure*

The lattice parameter values of HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ are calculated and noted in Table 1 after optimization of initiated design crystal geometry. Withal, the optimized structure is added in Fig. 1a–c which was taken after simulation GGA with PBE as the forthcoming functional from other method of GGA [24–26].

3.2 *Electronic Band Structure*

To scrutinize the electronic properties of HfO_2, the energy band and band gap structure are demonstrated in Fig. 2a, whereas the zero-point energy has been considered as the Fermi level. The band gap of HfO_2 puts on show accounting at 4.340 eV

Table 2 Band gap for HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$

Crystals/functional	GGA with PBE (eV)	Reference
HfO_2	4.340	4.340 [27]
$Hf_{0.88}Si_{0.12}O_2$	2.033	Newly predicted
$Hf_{0.88}Ge_{0.12}O_2$	1.686	Newly predicted
$Hf_{0.88}Sn_{0.12}O_2$	3.210	Newly predicted

as an indirect band gap having G symmetry point which is consistent overlapped with the reference results of 4.340 eV [27]. After 12% Si, 12% Ge and 12% Sn atom doping on HfO_2, the band gap has reduced at 2.033 eV, 1.686 and 3.210 eV for $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ as shown in Fig. 2b–d, but they have shown the direct band gap. By which it stands for that $Hf_{0.88}Ge_{0.12}O_2$ has a high capacity on HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ (Table 2).

3.3 The Density of States and Partial Density of State

The density of the state indicates the point of electronic band structures and the split of an orbital. The GGA with PBE method was used to calculate the density of states (DOS) of Hf, Si, Ge, Sn and O atoms of HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals. Form Fig. 3b–e, the sum of DOS for HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals consists of 6s2, 5p6, 4f14, 5d2 orbital for a Hf atom, 3s2, 2p6 orbital for Si atom, and 4s2,3p6, 3d10 orbital for Ge atom, 4d10, 5s2, 5p2 for Sn atom and 2s and 2p orbital for O atom.

Secondly, the conduction band found at the DOS of HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals at 0.00 to 5 eV. The DOS of the valance band is found at -0.1 to -5 electron/eV, while the DOS of the conduction band is recorded at about 10 to 20 electron/eV. To compare the s, p and d orbitals for both doping and undoped, the orbitals for $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ are much higher than HfO_2, and it can be said that the Si, Ge and Sn doping on HfO_2 have increased the DOS of any crystal showing in Fig. 3a–e.

4 Optical Properties

4.1 Reflectivity

Reflectivity is the measurement of reflected light from the surface area of the material relative to the amount of light incident on the material. In this investigation, the reflectivity of HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ is shown in Fig. 4. The initial reflectivity of HfO_2 has recorded 0.16, and with the increase of

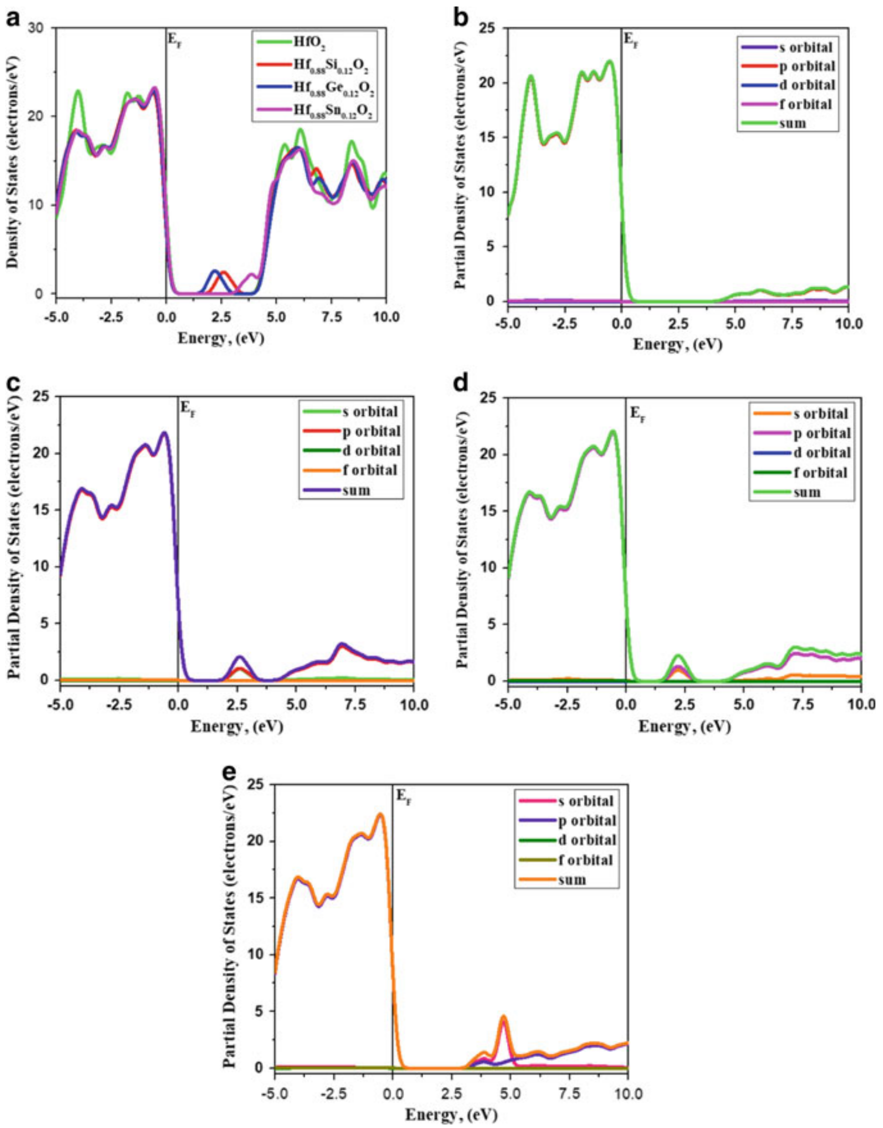


Fig. 3 **a** Total density of states for HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$, **b** partial total density of states for, **c** partial total density of states for $Hf_{0.88}Si_{0.12}O_2$, **d** comparison s orbital for $SrTiO_3$ and $Hf_{0.88}Ge_{0.12}O_2$, **e** partial total density of states for $Hf_{0.88}Sn_{0.12}O_2$

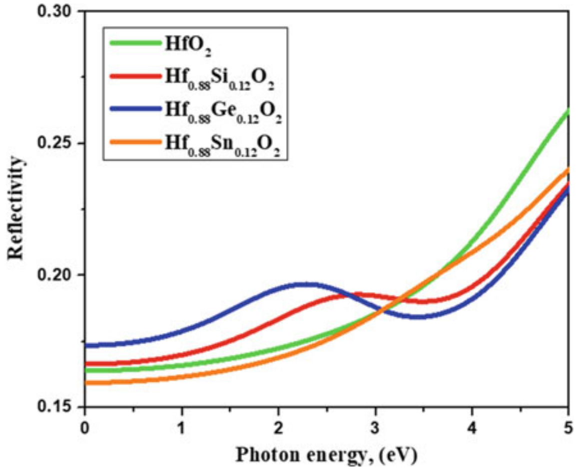


Fig. 4 Reflectivity of HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$

energy, it increases. The less reflectivity means more efficient quantum dots. After doping Si%, Ge% and Sn%, the initial reflectivity has recorded 0.17, 18 and 16 which decreases with the increase of energy and proved $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$.

4.2 Absorption

The absorption spectrum of a material depends on the scenery of the energy band gap which follows the indirect band gap usually absorbs more temperature than the direct band gap semiconductor device, because there are fewer phonons at low temperature. The absorption spectrum of $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ is higher than HfO_2. From Fig. 5, it is clear that with the increase of energy, the absorption of materials HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ increases but $Hf_{0.88}Ge_{0.12}O_2$ shows better value of absorption than HfO_2, $Hf_{0.88}Si_{0.12}O_2$, and $Hf_{0.88}Sn_{0.12}O_2$.

4.3 Refractive Index

The refractive index is one of the most important optical properties of a semiconductor material indicating the ability of passing light through the medium, as well as it is also related to the absorption of materials and band gap engineering. However, Fig. 6 exhibits the refractive index though the several of photon energy composing of the real part and the imaginary part for HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ showing a contrary pattern. With beginning of photon energy, the

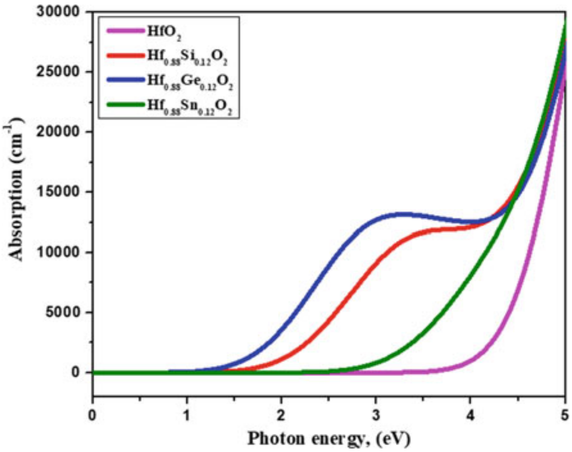


Fig. 5 Absorption of HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$

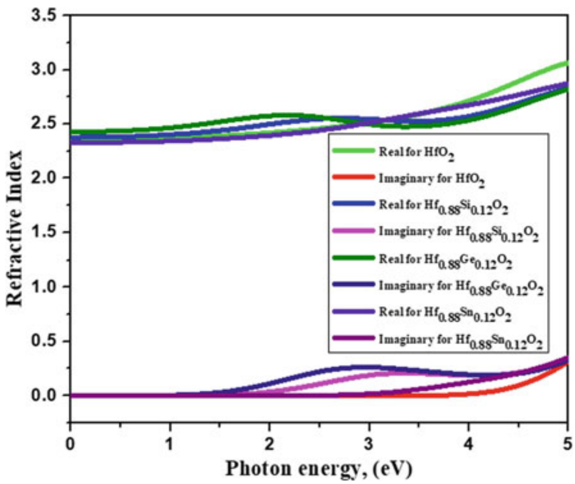


Fig. 6 Refractive index of HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$

refractive index for real part stays about 2.5 up to 1.0 eV and afterward they follow a constant pattern with slightly different values of refractive index, and the imaginary part is remained about 0.0 before 2.5 eV but it is fluctuated after this photon energy.

4.4 Dielectric Function

It is the most important factor for semiconductor devices, such as diode, MOSFETs and RRAM. Using the electric dipole approximation, the optical dielectric constant can be determined. In addition, it is composed of real and imaginary parts as the summation by following equation:

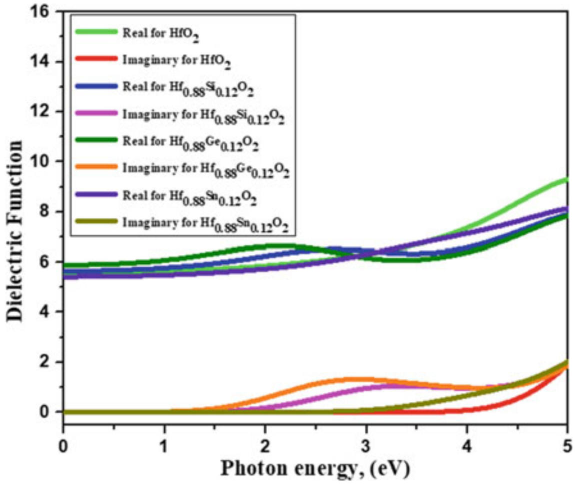


Fig. 7 Dielectric function of HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$

$$\varepsilon = \varepsilon_1(\omega) + i\varepsilon_2(\omega)$$

Here, ε_1 (ω) denotes the real part of dielectric constant, and the $\varepsilon_1(\omega)$ expresses the imaginary part as the dielectric loss factor. The probability of photon absorption for the band structure of any material is closely related to the imaginary portion of the dielectric constant, and also related the energy or charge capturing ability for transistor or battery. However, it is expressed the electric charge discharging ability of devices. Contrary, the real part of the dielectric constant is maintained the energy storage potential in the electric field. From Fig. 7, the real portion is always higher than the imaginary part within the energy at 1.0 to 1.8 eV so that it can be said that the HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ are highly energy storage materials within this energy. Afterward, it acts the opposite nature as the discharging materials in the range from 2.0 to 4.0 eV because the imaginary part shows higher value than real portion for HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$.

4.5 *Conductivity*

The conductivity can be explained in terms of band gap of electronic structure by the transition of electron from the valance band to the conduction band leaving the hole or charge. The cardinal fact of this forthcoming paragraph is related to the electron transition, while the larger value of conductivity mentions the higher ability of charge mobility and lower barrier. Figure 8 depicts the comparative study of the conductivity value of HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals. The conductivity values of both real and imaginary parts starting from almost zero at 0.0 eV photon energy. The real part of conductivity increased with a similar trend for HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ in the energy range

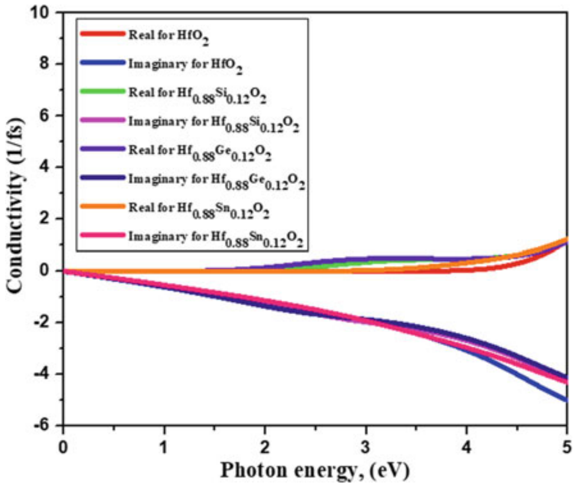


Fig. 8 Conductivity of HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$

from 0.0 eV to 5.0 eV and reached conductivity real peaked value 2.8 and 3.0 eV, but the conductivity value of $Hf_{0.88}Ge_{0.12}O_2$ within energy range 1.0–3.0 eV is higher than HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$. On the other hand, the imaginary part values of HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ are gradually declined after Fermi energy in the energy range from 3.0 eV, and reached conductivity imaginary peaked values −2.1 and −5.0.

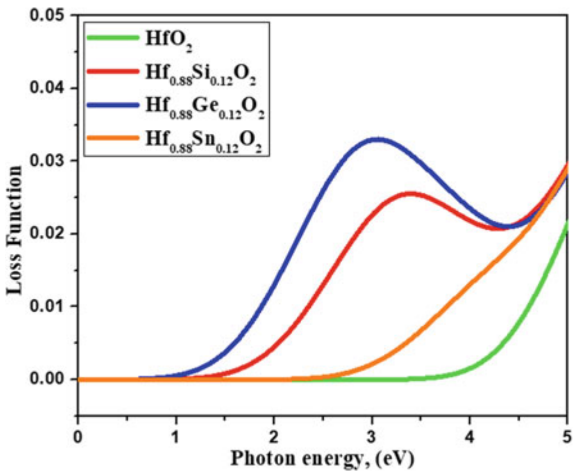


Fig. 9 Loss function of HfO_2, $Hf_{0.88}Si_{0.12}O_2$ $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$

4.6 Loss Function

The loss function is a central ingredient of optical properties. It is made by two regions on basis of photon energy, for instance, the lower photon energy division and higher photon energy division for crystal materials. There is a close relationship between loss function and dielectric function of solid materials. In the energy loss function, dielectric function reflects the response of a semiconductor to an external electromagnetic perturbation. The calculated exploration of loss function values for HfO_2, $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ illustrates in Fig. 9. It can be seen that loss function of $Hf_{0.88}Ge_{0.12}O_2$ is higher than HfO_2, $Hf_{0.88}Si_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$.

5 Conclusion

The metal oxide semiconductor, particularly the MOSFETs, in gate stack architecture is the cardinal materials for modern electronic device and technologies, while the HfO_2 crystal is the solution with doping the Si, Ge and Sn. The HfO_2 crystal conveys the band gap at 4.340 eV, and after doping Si%, Ge% and Sn%, the band gaps are reduced, while the band gap is recorded by 2.033 eV, 1.686 eV and 3.210 eV of $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ crystals, respectively. Secondly, the optical properties, due to Si, Ge and Sn doping, have changed. The optical dielectric function, absorption, reflectivity, loss function and conductivity of $Hf_{0.88}Si_{0.12}O_2$, $Hf_{0.88}Ge_{0.12}O_2$ and $Hf_{0.88}Sn_{0.12}O_2$ are greater than HfO_2. Finally, it can be said that the Si%, Ge% and Sn% doping in HfO_2 almost acted as a promising semiconductor material for MOSFETs and RRAM.

References

1. Colinge, J.-P.: Multiple-gate soi mosfets. Solid-State Electron. **48**, 897–905 (2004)
2. Lundstrom, M., Ren, Z.: Essential physics of carrier transport in nanoscale MOSFETs. IEEE Trans. Electron Devices **49**, 133–141 (2002)
3. Lee, H.C., Chen. P.S., Wu, T.Y., Chen, Y.S., Wang, C.C., Tzeng, P.J., Lin, C.H., Chen, F., Lien, C.H., Tsai, M.-J.: Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based RRAM. In: 2008 IEEE International Electron Devices Meeting, pp. 1–4 (2008)
4. Jagannathan, H.N., Narayanan, V., Brown, S.: Engineering high dielectric constant materials for band-edge CMOS applications. ECS Trans. **16**, 19 (2008)
5. Fantini, A., Wouters, D.J., Degraeve, R., Goux, L., Pantisano, L., Kar, G., Chen, Y.-Y., Govoreanu, B., Kittl, J.A., Altimime, L.: Intrinsic switching behavior in HfO_2 RRAM by fast electrical measurements on novel 2R test structures. In: 2012 4th IEEE International Memory Workshop, pp. 1–4 (2012)

6. Francescon, P., Cora, S., Cavedon, C., Scalchi, P., Reccanello, S., Colombo, F.: Use of a new type of radiochromic film, a new parallel-plate micro-chamber, MOSFETs, and TLD 800 microcubes in the dosimetry of small beams. Med. Phys. **25**, 503–511 (1998)
7. Stella, C., Laudani, M., Gaito, A., Nania, M.: Advantage of the use of an added driver source lead in discrete power MOSFETs. In: 2014 IEEE Applied Power Electronics Conference and Exposition-APEC 2014, pp. 2574–2581 (2014)
8. Picard, C., Brisset, C., Hoffmann, A., Charles, J.-P., Joffre, F.,Adams, L., Siedle, A.H.: Use of electrical stress and isochronal annealing on power MOSFETs in order to characterize the effects of 60Co irradiation. Microelectronics Reliab. **40**, 1647–1652 (2000)
9. Öttking, R., Kupke, S., Steve, Nadimi, E., Leitsmann, R., Lazarevic, F., Plänitz, P., Roll, G., Slesazeck, S., Trentzsch, M., Mikolajick, T.: Defect generation and activation processes in HfO_2 thin films: contributions to stress-induced leakage currents. Physica Status Solidi (A) **212**, 547–553 (2015)
10. Puglisi, F.M., Pavan, P., Larcher, L., Padovani, A.: Statistical analysis of random telegraph noise in HfO_2-based RRAM devices in LRS. Solid-State Electronics **113**, 132–137 (2015)
11. Pan, J.N.: Optoelectronic CMOS transistors: performance advantages for sub-7nm ULSI, RF ASIC, memories, and power MOSFETs. MRS Adv. **4**, 2585–2591 (2019)
12. Min, B., Devireddy, S.P., Çelik-Butler, Z., Wang, F., Zlotnicka, A., Tseng, H.-H., Tobin, P.J.: Low-frequency noise in submicrometer MOSFETs with HfO_2, HfO_2/Al_2O_3 and HfAlOx gate stacks. IEEE Trans. Electron Devices **51**, 1679–1687 (2004)
13. Fantini, A.: Intrinsic program instability in HfO_2 RRAM and consequences on program algorithms. In: 2015 IEEE International Electron Devices Meeting (IEDM), pp. 7.5.1–7.5.4 (2015)
14. Kumar, A., Chakma, U.: Developing the amazing photocatalyst of $ZnAg_2GeSe_4$, $ZnAg_2Ge_{0.93}Fe_{0.07}Se_4$ and $ZnAg_2Ge_{0.8}6Fe_{0.14}Se_4$ through the computational explorations by four DFT functionals. Heliyon **7**, e07467 (2021)
15. Al Mamun, A., Monsur Alam, M., Habib, A., Chakma, U., Sikder, M., Kumer, A.: Structural, electronic, optical properties and molecular dynamics study of WO_3 $W_{0.\,97}Ag_{0.\,03}O_3$ and $W_{0.94}Ag_{0.06}O_3$ photocatalyst by the first principle of DFT study. Egyptian J. Chem. **63** (2021)
16. Ali, M., Islam, M.J., Rafid, M., Jeetu, R.R., Roy, R., Chakma, U., Kumer, A.: The computational screening of structural, electronic, and optical properties for SiC, $Si_{0.94}Sn_{0.06}C$, and $Si_{0.88}Sn_{0.12}C$ lead-free photovoltaic inverters using DFT functional of first principle approach. Eurasian Chem. Commun. **3**, 327–338 (2021)
17. Hasan, M.M., Kumer, A., Chakma, U.: Theoretical investigation of doping effect of Fe for $SnWO_4$ in electronic structure and optical properties: DFT based first principle study. Adv. J. Chem.-Section A **03**, 639–644 (2020)
18. Howlader, D., Hossain M.S., Chakma, U., Kumer, A., Islam, M.J., Islam, M.T., Hossain, T., Islam, J.: Structural geometry, electronic structure, thermo-electronic and optical properties of $GaCuO_2$ and $GaCu_{0.94}Fe_{0.06}O_2$: a first principle approach of three DFT functionals. Mol. Simul. **44**, 1–12 (2021). https://doi.org/10.1080/08927022.2021.1977295
19. Islam, M.T., Kumer, A., Howlader, D., Chakma, K.B., Chakma, U.: Electronics structure and optical properties of $Mg(BiO_2)_4$ and $Mg(Bi_{0.91}Ge_{0.083}O_2)_4$: a first principle approach. Turk. Comput. Theor. Chem. **4**, 24–31 (2020)
20. Chakma, K.B., Kumer, A., Chakma, U., Howlader, D., Islam, M.T.: A theoretical investigation for electronics structure of $Mg(BiO_2)_2$ semiconductor using first principle approach. Int. J. New Chem. **7**, 247–255 (2020)
21. Sikder, M.A.M., Chakma, U., Kumer, A., Islam, M.J., Habib, A., Alam, M.M.: The exploration of structural, electronic and optical properties for MoS_2 and $Mo_{0.95}W_{0.05}S_2$ photocatalyst effort on wastewater treatment using DFT functional of first principle approach. Appl. J. Environ. Eng. Sci. **7**, 103–113 (2021)
22. Islam, M.T., Kumer, A., Chakma, U., Howlader, D.: A computational investigation of electronic structure and optical properties of $AlCuO_2$ and $AlCu_{0.96}Fe_{0.04}O_2$: a first principle approach. Orbital: Electron. J. Chem. **13** (2021)

23. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple. Phys. Rev. Lett. **77**, 3865 (1996)
24. Ali, M., Islam, M. J., Kumer, A., Hossain, M., Chakma, U., Howlader, D., Islam, M., Hossain, T.: Investigation of structural, electronic and optical properties of $Na_2InAgCl_6$, $K_2InAgCl\ _6$, and $Rb_2InAgCl\ _6$ Lead-free halide double Perovskites regarding with $Cs_2InAgCl_6$ Perovskites cell and a comparative study by DFT functionals. Mater. Res. **24** (2021)
25. Alam, M.M., Chakma, U., Kumer, A., Islam, M.S., Khandaker, N.Z., Hossain, T., Chakma, P., Uddin, J.: Investigation of structural, electronic, and optical properties of $CdAg_2GeSe_4$ and $HgAg2GeSe_4$ amazing photocatalyst regarding with $ZnAg_2GeSe_4$ and a comparative study by DFT functionals. SPAST Abstracts **1** (2021)
26. Howlader, D., Hossain, M.S., Chakma, U., Kumer, A., Islam, M.J., Islam, M.T., Hossain, T., Islam, J.: Structural geometry, electronic structure, thermo-electronic and optical properties of $GaCuO_2$ and $GaCu_{0.94}Fe_{0.06}O_2$: a first principle approach of three DFT functionals. Mol. Simul. **47**, 1411–1422 (2021)
27. Zhu, H., Tang, C., Fonseca, L.R.C., Ramprasad, R.: Recent progress in ab initio simulations of hafnia-based gate stacks. J. Mater. Sci. **47**, 7399–7416 (2012)

5G MIMO Antenna in Wireless Communication with Higher Efficiency and Return Loss

Afrin Binte Anwar, Rima Islam, Debalina Mollik, Prodip Kumar Saha Purnendu, Raja Rashidul Hasan, and Md. Abdur Rahman

Abstract With the advancement of wireless communications technology, 5G wireless mobile communication will be developed into a new generation. Multiple inputs multiple-output (MIMO) technology is predicted to be a useful element in the 5G communication area to enhance the performance that is yet unexplored in literature. This paper offers a suggested antenna model for minimizing return loss and a good efficiency factor for inset fed multiple-input-multiple-output (MIMO) microstrip patch antenna with a small size of mm is proposed for 38 GHz (Ka-band) which is in 5G frequency bands. Although many publications addressed this work, minimization of the system loss and increasing return loss, the gain is not yet found in the present literature. The proposed antenna consists of two antenna elements that are parallel to each other. Both antennas are constructed with the main substrate of Rogers RT 5880 and a superstate of polyimide film. The architecture of both patch antenna is 3.561 mm * 2.449 mm * 0.254 mm, return loss −66.009846 and −66.008594 dB, Voltage Standing Wave Ratio (VSWR) of 1.0010017 and 1.0010019, and gain 7.512 dB for both antennas. Considering all of the mentioned factors, the developed antenna is supposed to be suitable for or 5G communication technologies in the future.

Keywords CST · 5G · Gain · Ka-band · Radiation efficiency · Surface current · VSWR

1 Introduction

5G mobile communication is going to be commercialized soon, for the significant advancements in communication technology. However, many nations' 5G frequency bands have been implemented but return loss of 5G antennas could not minimize by the research and it is revealed by literature [1]. In the current situation, many countries

A. B. Anwar (✉) · R. Islam · D. Mollik · P. K. S. Purnendu · R. R. Hasan · Md. Abdur Rahman
American International University-Bangladesh, Kuratoli, Bangladesh
e-mail: afrinanwar144@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_19

throughout the world have transitioned entirely to 5G networks, whereas some lower-middle-income countries like Bangladesh, Bhutan, Maldives, and Pakistan and so on are going to make a plan to put out 5G network infrastructure in various stages within the following some years to respond to the rising demand of smartphones, intercellular networks as well [2].

Millimeter wave has been untapped spectrum (3–300 GHz) to meet the demand of the next generation [3]. Researchers are highly interested in 5G MIMO antennas because they have a large capacity and high data rate. MIMO was first in Bell laboratories from 1997 to 2002 under name Bell Laboratories Layer Space–Time (BLAST) [4]. Globally, mobile data use will increase a lot in 2025; per user average will increase up to 28 gigabytes in 2025. Per subscriber average will be in peak in the region of North America, per user it will be up to 50 gigabytes in 2025 [5]. As currently using spectrum is very crowded, 5G antenna technology is highly encouraged, and 5G allows the improved amount of data transmission [6]. 5G networks have been established in many high- and upper-middle-income countries, whereas many countries are yet aiming to plan infrastructure systems in stages over the next few years to accommodate the increased demand for smart phones and Internet-connected devices. The growth of the Internet of Things will be increased as well. Moving on to millimeter waves will provide new issues in antenna design for smartphones and base tower antenna. There is some issue about the frequencies in millimeter wave: these are propagation loss, signal attenuation, and path loss. Atmospheric disturbances are also among them due to its signal to interference noise ratio (SINR) being hampered [7]. To meet all of these massive numbers of people using the mobile network system, efficient antenna structures with small sizes, easy fabrication, and high performance will be required. The microstrip multiple-input-multiple-output (MIMO) technology is used for its ability to give multiplexing gain and diversity gain for better link quality and improved link capacity in wireless systems [8]. High mutual coupling between the elements is a problem in MIMO antenna it degrades the performance of a MIMO antenna system. This paper is to be arranged as Sect. 2 describes theory and methodology, Sect. 3 demonstrates the proposed antenna including design, Sect. 4 illustrates simulation results, a comparative study is shown in Sect. 5 and this paper will be completed by the conclusion. Necessary references have been mentioned after the conclusion.

2 Theory and Methodology

2.1 Equation for Antenna Designing

For enhanced performance, a multi-slotted insert feed has been designed for nicrostrip MIMO antenna to work it on at 38 GHz. In free space, the geometric view is given of the designed antenna in Fig. 1.

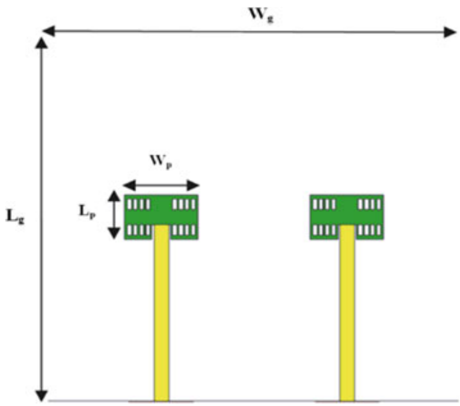


Fig. 1 Approached microstrip patch MIMO antenna's structure (in free space).

In a simulation environment, the maximum and minimum frequencies are first determined. The maximum and minimum frequency ranges are 40 GHz and 35 GHz, respectively. Then as a substrate, Rogers RT5880 (lossy) was selected which had 2.2 permittivity and polyimide is chosen as a superstate purpose (Fig. 2).

The frequency of 38 GHz has been chosen as the operating frequency. Width and length of the patch are computed using the calculations below.

$$W = \frac{c_0}{2 f_r \sqrt{\frac{\varepsilon_r + 1}{2}}} \tag{1}$$

The specified frequency width is calculated by using Eq. (1). As bandwidth and gain are proportional to width, further modification is needed. If W aperture area will increase mean gain will increase and the fringing field will increase mean more radiation will occur. However, in design, antenna's width cannot be increased because there are some physical limitations for fitting it in some particular devices.

Using this equation, effective permittivity $\varepsilon_{\text{reff}}$ is calculated.

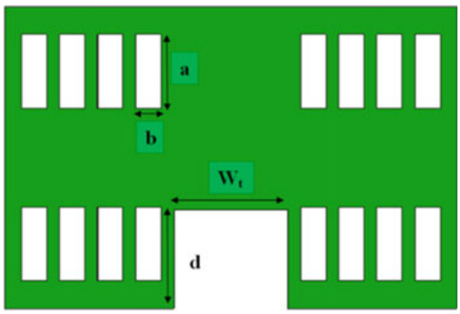


Fig. 2 Microstrip patch antenna's structure with slot size, depth, and feedline width

$$\varepsilon_{\text{reff}} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 2}{2}\left(1 + 12\frac{h}{w}\right)^{-0.5} \tag{2}$$

Effective length, L_{eff}, is calculated using this equation using previously determined effective permittivity value.

$$L_{\text{eff}} = \frac{c_0}{2 f_r \sqrt{\varepsilon_{\text{reff}}}} \tag{3}$$

where L_{eff} = Effective length.

And then using the below equation, length extension ΔL due to fringing field is calculated.

$$\Delta L = 0.412 \frac{\left(\frac{w}{h} + 0.264\right)(\varepsilon_{\text{reff}} + 0.3)}{(\varepsilon_{\text{reff}} - 0.258)\left(\frac{w}{h} + 0.813\right)} \tag{4}$$

Using L_{eff} and ΔL, actual length is calculated.

$$L = L_{\text{eff}} - 2\Delta L \tag{5}$$

However, mutual coupling is a problem in MIMO antenna. Several techniques are there to reduce mutual coupling and those are the spacing between two elements should be greater than λ/4, placing two elements perpendicular two each other, using parasitic element and so on [9]. However, there are only two elements as result it is possible to keep the space between the elements are very large compare to λ/4.

After getting all parameters considering all the above equations, the MIMO antenna is designed in CST. Then cutting multiple slots and keeping a proper between the two-element, the operating frequency radiation pattern and mutual coupling are avoided [10].

2.2 Mutual Coupling Analysis of MIMO Antenna

The definition of mutual coupling is when one antenna transmits energy and that energy is absorbed by the nearest antenna. Usually, spacing between antennas is much little in mobile applications what affect the performance of the antenna. Due to channel correlation, the system's performance usually deteriorates in MIMO systems. The reason for this is the mutual coupling which makes changes in the distribution of current. As a result, there are distortions of the radiation pattern in every antenna element. On the other hand, radiation pattern distortion increases diversity gain and eventually then shows better MIMO antenna performance. Mostly, surface and space waves are responsible for mutual coupling. The MC_{ij} empirical mutual coupling model is represented as

$$\mathrm{MC}_{ij} = \exp\left(-\frac{2d_{ij}}{\lambda}(\alpha + j\pi)\right), i \neq j \tag{6}$$

$$\mathrm{MC}_{ij} = 1 - \frac{1}{N}\sum_{i}\sum_{i \neq j}\mathrm{MC}_{ij} \tag{7}$$

Here,

MC = Mutual coupling
$d_{ij} = d_{ij}$ is the distance in the middle of i-th and j-th antenna elements
N = the number of array elements
α = controlling parameter of the coupling level.

The given structure gives ideas about reduction techniques of mutual coupling. Here, bandwidth, efficiency, isolation level, diversity gain, substrate materials, and envelop correlation efficiency are compared with the system performances of MIMO antenna [9, 11].

2.3 Reduction Techniques for Mutual Coupling

See Fig. 3.

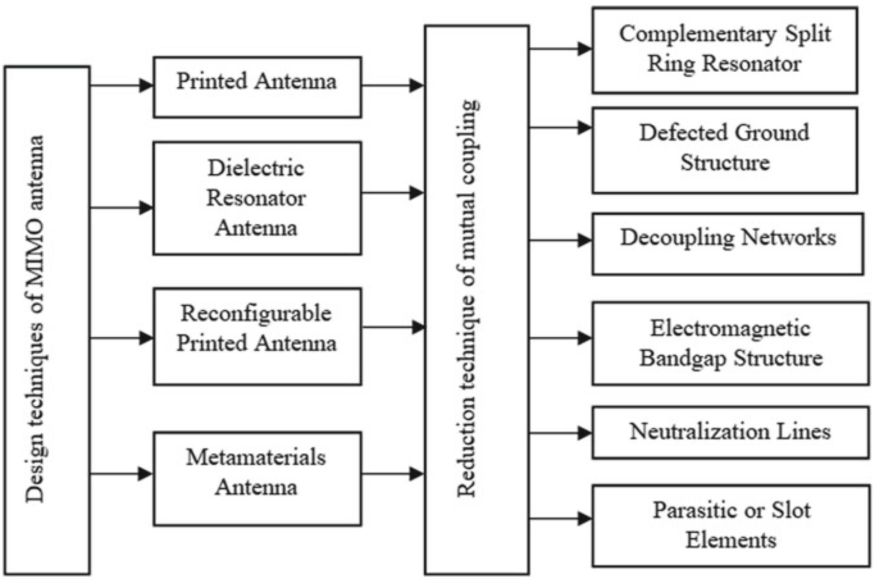


Fig. 3 Different approaches of the MIMO antenna's mutual coupling reduction techniques [9]

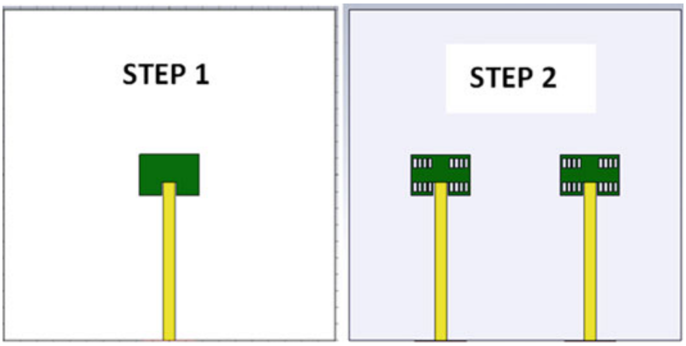


Fig. 4 Step 1 is the primary architecture of antenna and step 2 is the final view of antenna

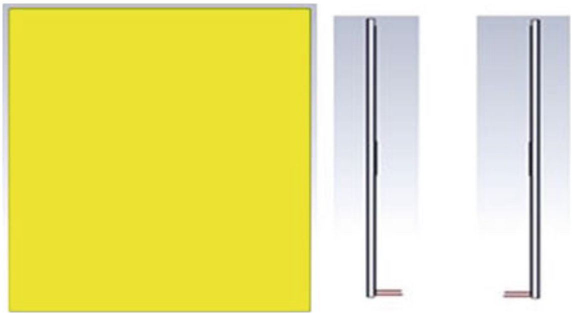


Fig. 5 Rear view of antenna and side views of antenna

3 Antenna Design

CST studio 2016 version was used to determine the simulated design, as well as its geometry and various viewpoints. The designed microstrip patch antenna is depicted in the front, rear, sideways, top, and bottom views (Figs. 4 and 5).

The following are the parameters that have been utilized during the design of the suggested microstrip MIMO patch antenna (Table 1):

4 Simulation and Analysis

In this section, all simulation results have been obtained and discussed. This paper has been contributed to innovation in research works for communication field. New contributions have been stated below:

- This paper has higher return loss which is −66.008594 dB.
- High efficiency and high gain.
- Mutual coupling is −55.822797 dB which is considered good.

Table 1 Parameters of proposed antenna's model

Model's symbols	Model's parameters	Dimension (mm)
W_p	Patch's width	3.561
L_p	Patch's length	2.449
W	Ground/substrate plane's width	20
L	Ground/substrate plane's length	20
a	Slot width	0.2
b	Slot length	0.3
Sh	Height of substrate	0.254
f_t	Feedline width	0.695
h_t	Ground thickness	0.035
GW	Distance between patch and feedline	0.1
d	Depth	0.8

4.1 S-Parameter

Computer simulation software CST 2016 version is used for measuring the S-parameter of this MIMO antenna. From the simulated result, port 1 and port 2 also have a similar bandwidth of 37.5–38.936 GHz for s11 < −10 dB. Thus, the MIMO has been satisfied the impedance matching criteria. Mutual coupling less than −15 dB is considered good for MIMO antennas. In 2 elements MIMO, $S_{2,1}$ and $S_{1,2}$ represent mutual coupling means transmission cosecant between two elements. In the simulated results, $S_{1,1}$ −66.009846 dB and $S_{2,2}$ −66.008594 dB, $S_{2,1}$ is −55.822797 dB and $S_{1,2}$ is −55.82278 dB at resonant frequency 38.77 GHz which is very good for antenna performance (Figs. 6 and 7).

4.2 VSWR

The ideal value of VSWR is 1 where maximum power is transferred. The obtained VSWR of the preferred antenna model at port 1 is 1.0010017 (VSWR1), and in port 2 is 1.0010019 (VSWR2) at the operating frequency of 38.225 GHz, as shown in Figs. 8 and 9 which is good for power delivery.

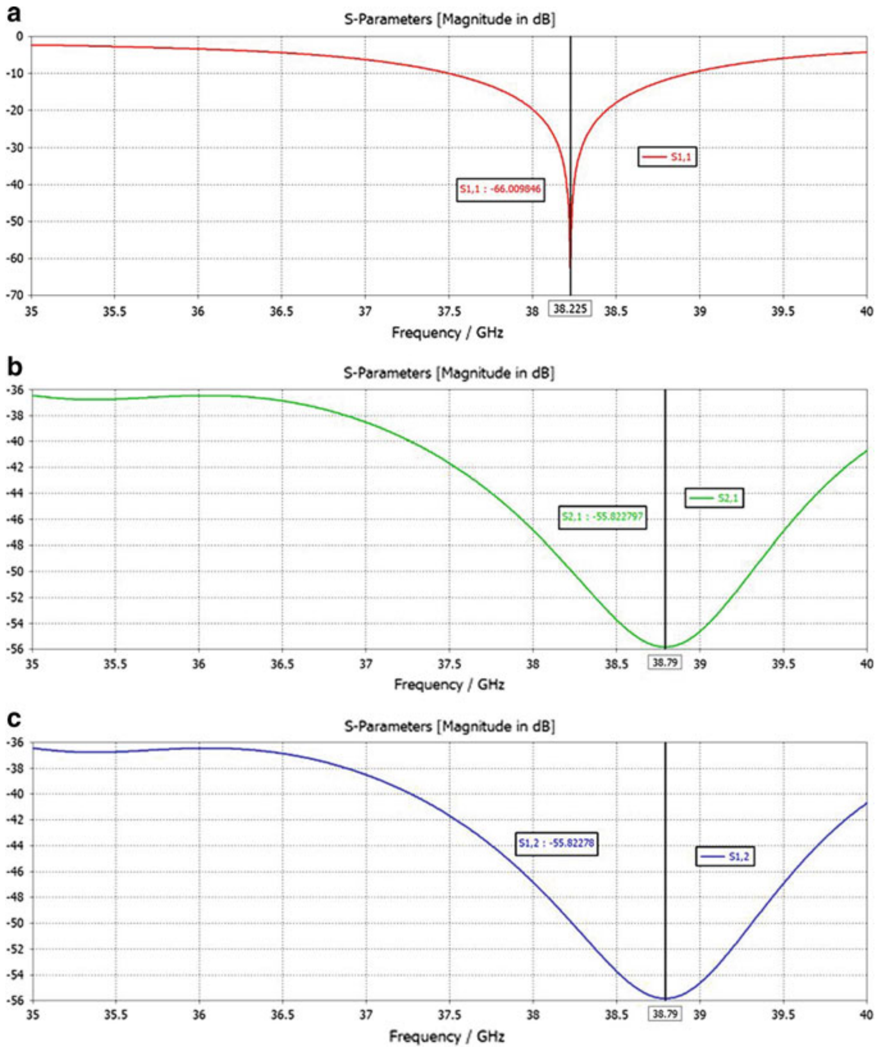


Fig. 6 **a** and **d** are the antenna's return loss for free space, whereas **b** and **c** are the reverse isolation and the insertion loss due to passive device case

4.3 Result Analysis of Gain and Directivity

Both of gain and directivity are crucial parameters which are considered during evaluation of an antenna's performance. Antenna gain is a performance parameter that is measured against a reference source. 7.512 dB is the gain of the designed antenna (Figs. 10, 11, 12, 13, and 14).

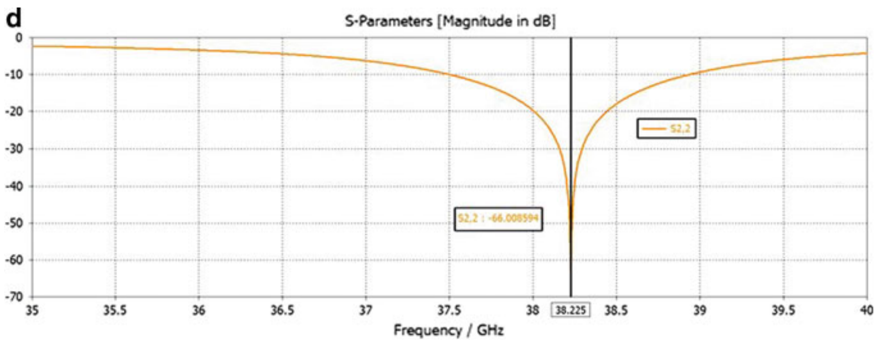


Fig. 6 (continued)

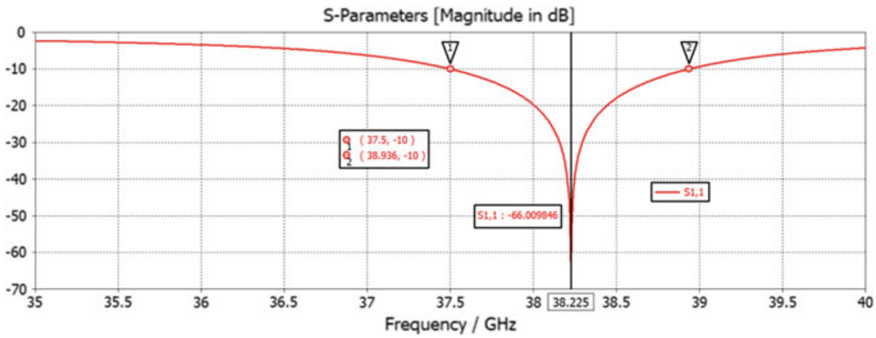


Fig. 7 Bandwidth of the suggested antenna design

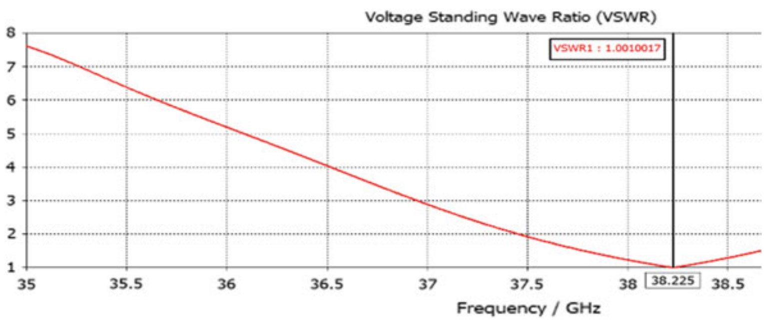


Fig. 8 VSWR versus the frequency graph in port 1 (shown at the free space)

4.4 Overall Efficiency and Radiation Efficiency

This formula is used to determine antenna efficiency (Fig. 15).

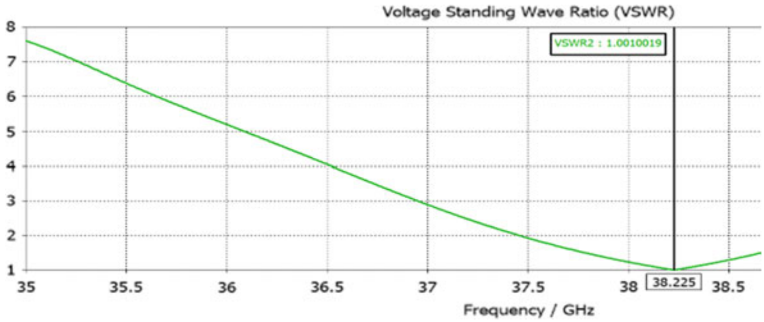


Fig. 9 VSWR versus the frequency in port 2 (shown at the free space)

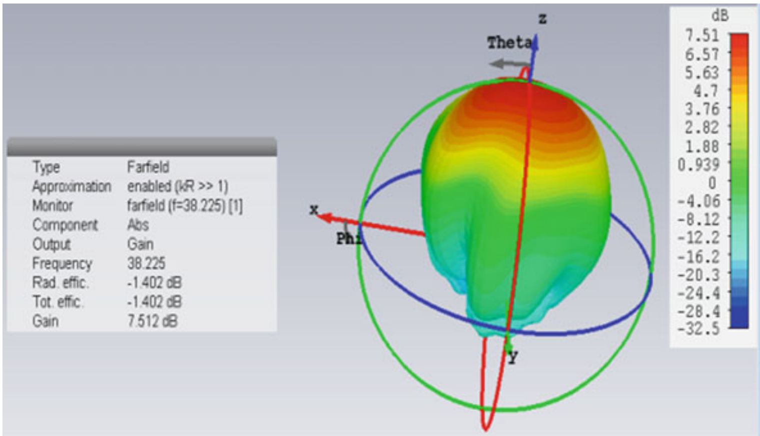


Fig. 10 3D view of gain of the designed antenna for frequency 38.225 GHz (in free space)

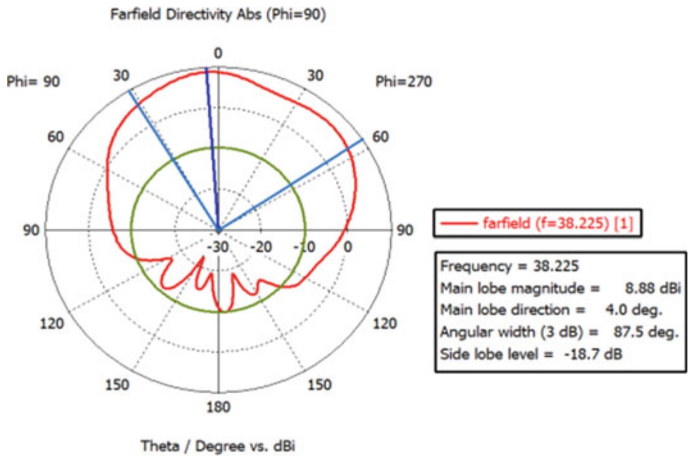


Fig. 11 For frequency 38.225 GHz, a 2D polar chart of gain for the antenna

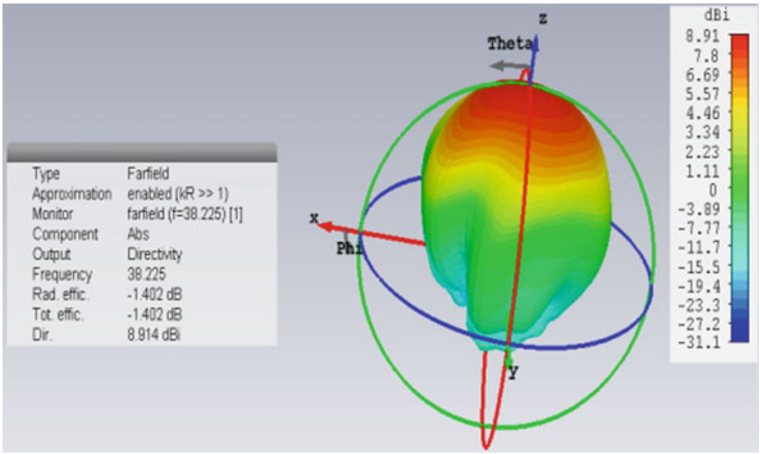


Fig. 12 Radiation pattern in three dimensions at frequency 38.225 GHz (directivity) (in free space)

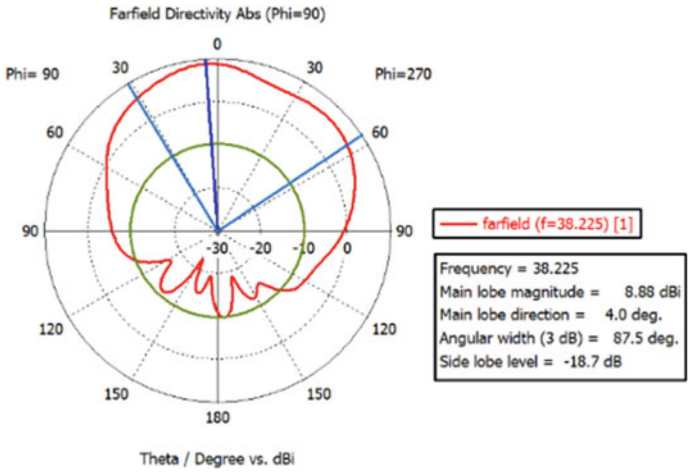


Fig. 13 Radiation pattern (directivity) 2D far field directivity (in free space) for frequency 38.225 GHz

$$\text{Efficiency} = \left[\text{Gain}/\text{Directivity}\right] \times 100\%$$

4.5 *Surface Current and Smith Chart*

The surface current is 406.4 A/m. In antenna analysis, as a function of frequency, the Smith chart shows antenna impedance. Using this chart, proper modification can

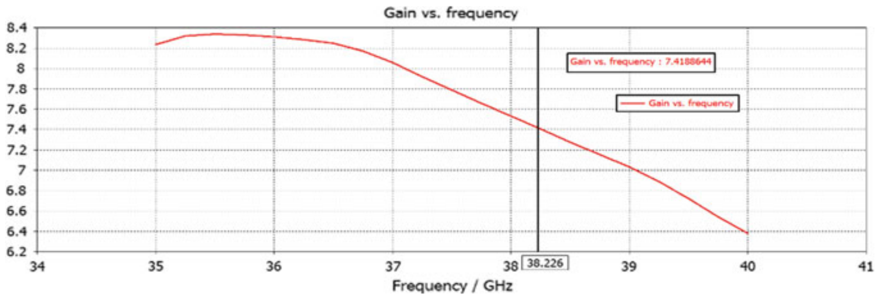


Fig. 14 Gain in opposition to frequency graph (shown in free space)

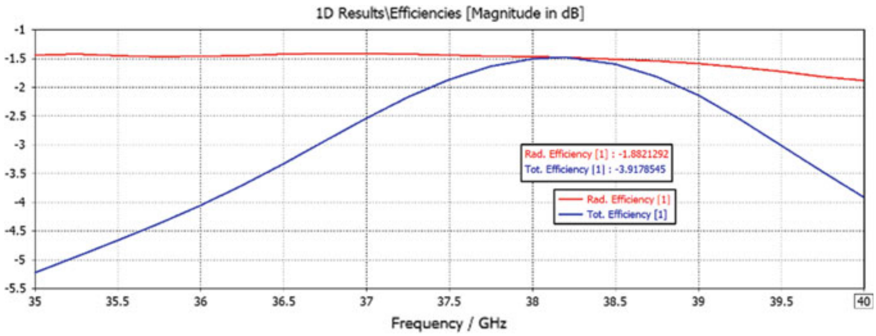


Fig. 15 Radiation (red graph) and total (blue graph) efficiency curve (shown in free space)

be made if a proper match is not obtained in a particular frequency. In this proposed antenna at resonant frequency 38.225 GHz, the input impedance had been gotten all most matched to the ideal 50 Ω coaxial feed line (Figs. 16 and 17; Table 2).

5 Comparative Study

Various antenna characteristics, such as the S-parameter ($S_{1,1}$, $S_{2,2,}$ $S_{1,2}$, $S_{2,1}$), Voltage Standing Wave Ratio (VSWR), gain, directivity, bandwidth, efficiency, and resonant frequency, were investigated in this study and have been identified in many recent research publications. After doing a simulation on CST at 38 GHz, the resonant frequencies were discovered to be 38.225 and 38.77 GHz. Better $S_{1,1}$ and $S_{2,2}$ parameters were found with good mutual coupling when it had been analyzed with the different references and articles given in Table 3. 5G communication system with optimal gain, efficiency, and bandwidth has been received. In the table, different parameters and simulation results which are shown in the last row were analyzed with many previous research works. Table 3 demonstrates that, when compared to models produced in the previous experimentation, the greater value of $S_{1,1}$ and $S_{2,2}$

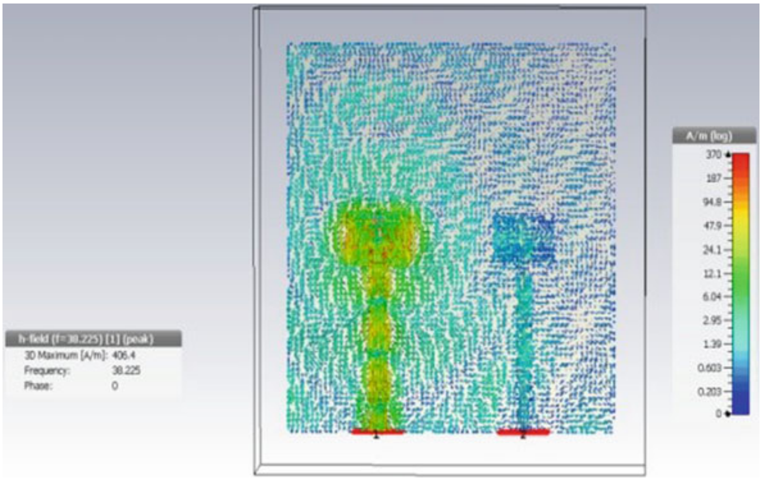


Fig. 16 Surface current of recommended antenna (shown in free space)

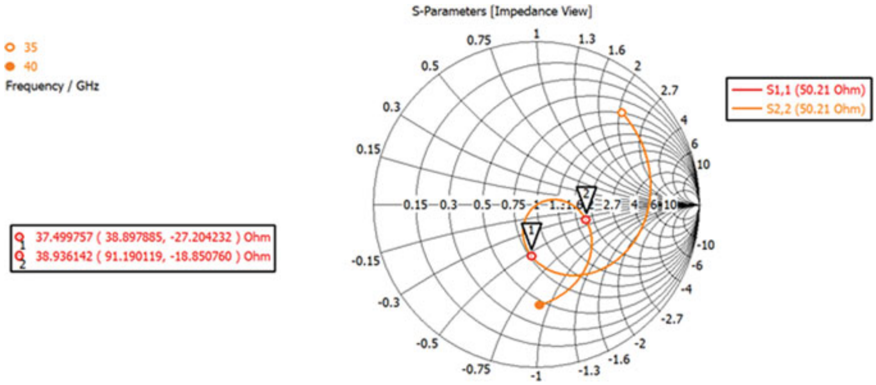


Fig. 17 Smith chart diagram which has shown the impedance view of the antenna

was obtained; hence, the proposed model absorbs more power into the antenna due to the high return loss. The $S_{1,1}$ and $S_{2,2}$ parameters are the most important for high-speed procedures. The suggested antenna's high bandwidth will be advantageous for speedy data transfer in a 5G wireless communication.

Table 4 also provides a contrastive experimentation of the proposed antenna models. This table briefly discussed antenna architecture and substrate material and compared with the earlier mentioned antenna's architecture in different research works. From the correlation, it can be observed that the suggested antenna has improved simulation results than the others antenna due to the compact architecture. Furthermore, different substrate materials have produced different results in recent studies. For any antenna-related application, technology of new generation

Table 2 Brief overview of the obtained parameters of the recommended antenna

Parameters of antenna	Obtained values
$S_{1,1}$	−66.009846 dB
$S_{2,1}$	−55.822797 dB
$S_{1,2}$	−55.82278 dB
$S_{2,2}$	−66.008594 dB
Bandwidth	1.436 GHz
Gain [38.225]	7.512 dB
VSWR1, VSWR2	1.0010017, 1.0010019
Efficiency [38.225]	84.27%
Radiation efficiency [38.225]	−1.402 dB
Directivity [38.225]	8.914 dBi
Surface current [38.225]	406.4 A/m

includes higher gain, efficiency, and smaller size. As a result of the comparative analyses shown above, it may be concluded that the recommended antenna could be a preferable alternative for wireless communication than the previous antenna's architecture.

6 Conclusion

The proposed work has led us to conclude that the proposed MIMO microstrip patch antenna is compatible with the upcoming 5G technology. Operating frequency of the suggested antenna is 38.225 GHz for both antennas which have a gain of 7.512 dB, the efficiency of antenna is 84.27%. Radiation efficiency at 38.225 GHz frequency is −1.402 dB. This antenna has total efficiency −3.91 dB. The radiation pattern of this antenna has showed that it will work only in one direction. Hence, the simulation results have been showed that this antenna has better gain, return loss, and efficiency at the operating frequency 38.225 GHz. The results indicate that the proposed antenna can withstand the higher operating frequency in 5G communications. The design of the antenna is quite simple and kept appropriate spacing between two antennas ensuring proper mutual coupling. Future studies should target the applications like biomedical, industry, and other scientific fields. This work can be provided as a reference for future 5G applications research.

Table 3 Comparative statement with others research work

S-parameter	Resonant frequency (in GHz)	Gain	Bandwidth in (GHz)	Efficiency (%)	Reference paper
$S_{1,1} = -24.35$ dB	38	–	1.021	–	[12]
$S_{1,1} = -44.71$ dB	28.5	18.5 dBi	1.93	–	[13]
$S_{1,1} = -22$ dB	28	–	2.55	78	[14]
$S_{1,1} = -15$ dB	38	–	2.1	76	
$S_{1,1} = -15.5$ dB	38	6.9	1.94	93.5	[15]
$S_{1,1} = -12$ dB	54	7.4	2.05	82.7	
$S_{1,1} = -24.80$ dB	28.73	13.2dBi	7.48	63.42	[16]
$S_{1,1} = -\ 66.009$ dB	38.225	7.512 dB	1.436	84.27	This work
$S_{2,2} = -\ 66.008$ dB	38.225	7.512 dB	1.436	84.27	This work

Table 4 Comparison of obtained results with others

Antenna's architecture (length × width × substrate height)	Description of materials (substrate)	Reference paper
8 × 5.9 × 1.6	FR4	[12]
116.2 × 196.88 × 1.575	RT duroid 5880	[13]
14 × 12 × 0.38	Rogers RT/ duroid 5880	[14]
6 × 6.25 × 0.578	Rogers RT5880 (lossy)	[15]
25 × 35 × 0.8	FR4	[16]
3.561 × 2.449 × 0.254	Rogers RT5880 (lossy)	This work

References

1. Luo, N., He, Y., Zhang, L., Wong, S.-W., Li, C., Huang, Y.: A differential broadband dual-polarized base station antenna element for 4G And 5G applications. In: 2019 Computing, Communications and IoT Applications (ComComAp) (2019)
2. Alsharif, M.H., Nordin, R.: Evolution towards fifth-generation (5G) wireless networks: current trends and challenges in the deployment of millimetre wave, massive MIMO, and small cells. Telecommun. Syst. **64**(4), 617–637 (2016)
3. Imran, D., Farooqi, M.M., Khattak, M.I., Ullah, Z., Khan, M.I., Khattak, M.A., Dar, H.: Millimeter wave microstrip patch antenna for 5G mobile communication. In: 2018 International Conference on Engineering and Emerging Technologies (ICEET) (2018)
4. Malathi, A.C.J., Thiripurasundari, D.: Review on isolation techniques in MIMO antenna systems. Indian J. Sci. Technol. **9**, 35 (2016). https://doi.org/10.17485/ijst/2016/v9i35/96704
5. Mobile Data Usage Per User Worldwide by Region 2019–2025|Statista, Statista (2021) (Online). Available: https://www.statista.com/statistics/489169/canada-united-states-average-data-usage-user-per-month/. Accessed 19 Jun 2021
6. Alsharif, M.H., Nordin, R.: Evolution towards fifth-generation (5G) wireless networks: current trends and challenges in the deployment of millimetre wave, massive MIMO, and small cells. Telecommun. Syst. **64**(4), 617–637 (2017)
7. Murugan, S.: Compact square patch antenna for 5G communication. In: 2nd International Conference on Data, Engineering and Applications (IDEA), pp. 1–3, IEEE (2020)
8. Zheng, L., Tse, C.: Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. IEEE Trans. Inf. Theory **49**, 1073–1096 (2003)
9. Nadeem, I., Choi, D.-Y.: Study on mutual coupling reduction technique for MIMO antennas. IEEE Access **7**, 563–586 (2019)
10. Balanis, C.A.: Antenna Theory: Analysis and Design. Wiley (2016)
11. Zhang, H., Wang, Z., Yu, J., Huang, J.: A compact MIMO antenna for wireless communication. IEEE Antennas Propag. Mag. **50**(6), 104–107 (2008). https://doi.org/10.1109/MAP.2008.4768932
12. Seker, C., Güneşer, M.: A single band antenna design for future millimeter wave wireless communication at 38 GHz. Eur. J. Eng. Formal Sci. **2**, 34–38 (2018). https://doi.org/10.2478/ejef-2018-0009
13. Rahayu, Y., Mustofa, I.R.: Design of 2×2 MIMO microstrip antenna rectangular patch array for 5G wireless communication network. In: 2017 Progress in Electromagnetics Research Symposium (PIERS-FALL), pp. 2679–2683 (2017). https://doi.org/10.1109/PIERS-FALL.2017.8293591
14. Hasan, M.N., Bashir, S., Chu, S.: Dual band omnidirectional millimeter wave antenna for 5G communications. J. Electromagnetic Waves Appl. **33**, 1–10 (2019). https://doi.org/10.1080/09205071.2019.1617790

15. Imran, D., Farooqi, M., Khattak, M., Ullah, Z., Khan, I.M., Khattak, M., Dar, H.: Millimeter Wave MicroStrip Patch Antenna for 5G Mobile Communication (2019)
16. Jose, M., Sankararajan, R., Sreeja, B.S., Kumar, P.: Laser micro-machined 28 GHz broad band single feed microstrip antenna for 5G mm-wave applications. Circuit World **46**, 6–12 (2019). https://doi.org/10.1108/CW-06-2019-0053

Designing of a Charging Capacity Model (CCM) for Electric Vehicles and Easy-Bikes

Shafquat Yasar Aurko, Riyad Salehdin, Faiaz Allahma Rafi, and K. Habibul Kabir

Abstract On the verge of the 4th Industrial Revolution, the growth of electric vehicles (EVs)(e.g., Easy-bikes) all over the world, especially in developing countries, leads to grid overloading due to non-coordinated charging of large number of EVs. Coordinated charging of EVs is expected to mitigate this problem. This paper proposes a variant model of Coordinated charging, i.e., Charging Capacity Model (CCM), which sets a limit on the number of EVs allowed to be charged at different hours of the day. The proposed CCM uses a modified Particle Swarm Optimization (PSO) algorithm. For implementation, Internet of Things (IoT) and data analytics are also used. The proposed model is simulated for the case of Bangladesh, with a specific type of EV called the Easy-bike, which is a popular three-wheeler vehicle in the country. This model achieves peak shaving and valley filling for the EV load curve and calculates hourly charging capacities for each charging station for the average load curve of eleven months. These results enhance the policymakers to mitigate any overloading problem due to EV integration.

Keywords PSO · Particle Swarm Optimization · Electric vehicles · Easy-bikes · EV · Three-wheeler · Coordinated charging · IoT · Data analysis · Communication · Networking

---

S. Yasar Aurko · R. Salehdin · F. Allahma Rafi · K. Habibul Kabir (✉)
Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Dhaka 1704, Bangladesh
e-mail: habib@iut-dhaka.edu

S. Yasar Aurko
e-mail: shafquatyasar@iut-dhaka.edu

R. Salehdin
e-mail: riyadsalehdin@iut-dhaka.edu

F. Allahma Rafi
e-mail: faiazallahma@iut-dhaka.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_20

1 Introduction

In the wake of decreasing fossil fuel reserves, the emergence of electric vehicles (EVs) as a replacement for fossil fuel-driven conventional vehicles is one of the dominant features of the 4th Industrial Revolution. EVs have a very low carbon footprint and produce no tailpipe emissions. As conventional vehicles emit greenhouse gas, the proliferation of EVs would greatly help in reducing global warming. Also EVs are 2–4 times more efficient compared to conventional fuel-based vehicles [1]. All these are making EVs a great demand on the market. There are 10 M registered EVs in the world by 2020, a year which has seen increase of EV sales by 41%. However, the rapid increase of electric vehicles comes at a great cost—it needs a well-developed power generation and distribution system as well as plentiful and sophisticated charging infrastructures. Unfortunately, many countries do not have the charging infrastructure and the suitable power grid needed to support the simultaneous charging of large numbers of EVs. Such unplanned and non-coordinated charging results in increasing the chances of grid overloading, which in turn, inadvertently leads to load shedding. The most feasible solution to this problem is a coordinated charging scheme, which prevents overloading in the power grid due to EV charging. This paper proposes the Charging Capacity Model (CCM). The proposed model accounts for the number of EVs present all over the country. An optimized EV load curve is generated, using a modified Particle Swarm Optimization (PSO) algorithm to achieve peak shaving and valley filling. It prevents the total load from surpassing the generation capacity and uses that to set limits on the number of EVs that can be charged at a given time, which is taken as one hour in this paper. From there, the number of charging stations required throughout the country is calculated, and the number of EVs allowed at each charging station at each hour is obtained. Using Internet of Things (IoT) devices, through data analysis and cloud computing, the Distribution System Operator (DSO) can monitor the situation on ground and can notify both the charging stations as well as the EVs. The system envisioned by the CCM is shown in Fig. 1.

There are two main objectives for this paper. The first objective is to find the charging capacities for each time interval throughout 24 h. The second objective is to find the optimized number of charging stations and using that information, to find the number of EVs that can be charged per charging station at an hourly basis. In this paper, this model has been demonstrated for the case of Bangladesh as the test zone. A specific type of EV, the electric three-wheeler called Easy-bikes (EB), is considered in this model. This is because it is the most widespread EV type in Bangladesh at present, and thus, the EV with the most significant effect on the power grid of the country. But unfortunately for Bangladesh, the charging infrastructure and the power sector are not developing as fast [2]. Thus, this model is well suited for implementation in the scenario of Bangladesh. For the given data of the electric grid and the EBs in Bangladesh, this model achieves peak shaving and valley filling for the load curve of EB charging and suggests an optimized load curve, which is then used to find the charging capacities at different hours of the day. This includes the total number of EBs allowed to be charged at each hour throughout the country,

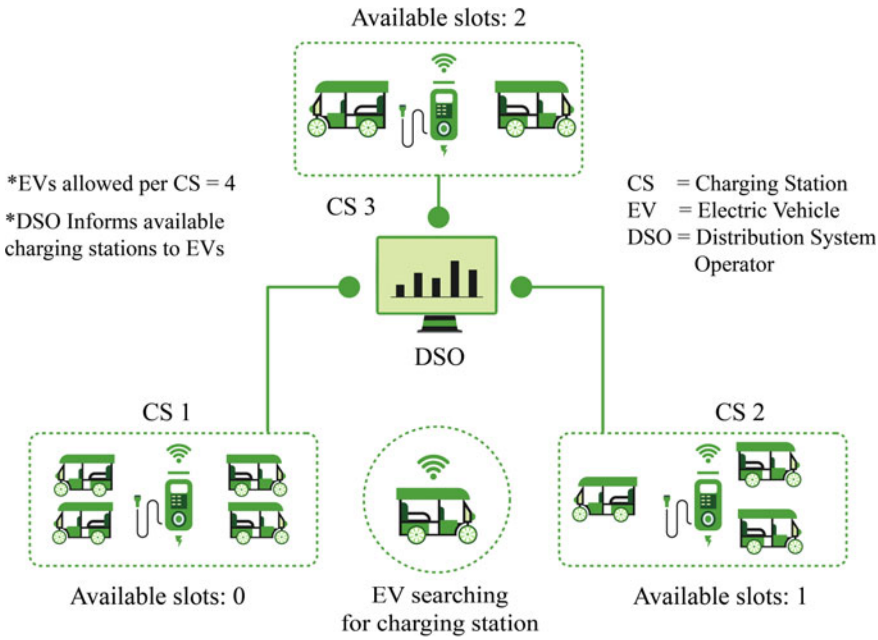


Fig. 1 System envision by the proposed model

as well as the number of EBs that can be charged at each charging station at an hourly basis, which has been demonstrated in this paper for the total load of eleven months of 2021. This provides a total load curve with reduced peaks and raised valleys, minimizing the chances of grid overloading.

2 Literature Review and Related Works

Easy-bikes (EB) are the most widespread EVs in Bangladesh, with currently 1 M of these all over the country [3, 4]. Their utility as a ride-sharing vehicle with relatively low cost has led to their popularity even with little promotion from the authorities. At present, Easy-bikes are responsible for providing primary transportation for over 250 M people [5], which indicates that they have a great effect on the grid compared to other types of EVs. Despite this, there are only 14 standardized EV charging stations operating in Bangladesh. These have a total capacity of 278 kW [6], whereas the charging of the 1 M Easy-bikes in the country consumes a total of 450 MW on a daily basis [5]. Most of these are charged using household connections [7], from locally set up charging shops [8]. Due to legal loopholes, licenses are not issued to Easy-bikes by the government of Bangladesh [9], leading to lack of regulation, inefficient design, and usage by inexperienced drivers, often leading to accidents.

With problems mounting, the Bangladeshi government officials are finally starting to take steps for legalizing and thus regulating Easy-bikes. BEPRC has started preparing guidelines for Easy-bikes [10], and licenses are to be provided to Easy-bike drivers to bring them under regulation [11]. In such a situation, the proposed CCM is well suited for application and acceptance in Bangladesh.

Non-coordinated charging of EVs is linked to many negative impacts, among which the power overloading problem is clearly the most significant problem with very tangible effects. Most of the research works on this subject involve load shifting through real-time scheduling with the use of heuristic algorithms, utilizing user data and preferences [12–14]. However, controlling and scheduling the EVs directly is difficult to implement in countries with overpopulation and huge numbers of EVs, especially when many of these take away the freedom of the EV users in the scheduling, making it a very automated process.

The proposed CCM only sets limits on the charging capacities at different hours of the day. In contrast to the noted works, this model is far easier and cheaper to implement as it does not involve real-time scheduling. CCM is more practically implementable as it is easier to control the charging capacity of charging stations than the behavior of EV users. PSO has been chosen for the optimization of EV load curve. This is because it has some distinct advantages in comparison with other optimization algorithms which include ease of operation, fewer variables to work with, faster convergence time, and also its global optimum can be further enhanced [15]. The PSO algorithm has been modified for optimizing the EV load in this model.

The next section describes the proposed CCM model in detail.

3 Proposed Charging Capacity Model (CCM)

The general mathematical model is developed in this section. It can be applied for all types of EVs. In this paper, only Easy-bikes are being considered, as this is the most widespread EV type in Bangladesh, which is the test case. The operation of the CCM model consists of five steps which are described below.

3.1 Generating the Electric Vehicle Load Curve

The first step to the CCM operation is to obtain the electric vehicle load curve data. For this, first the average power consumed by an individual electric vehicle [16], P_{EV_i}, is calculated. This is shown in Eq. (1),

$$P_{\mathrm{EV}_i} = (1 - \mathrm{SOC}_i)\frac{E_{\max_i}}{\mathrm{CT}_i}, \tag{1}$$

where i is the identity of the Electric Vehicles ($i = 1, 2, 3, \ldots, N$), SOC_i is the state of charge , $E_{\max_i}$ is the maximum energy that can be stored in the battery of EV_i (depends on the type of EV), and CT_i is the Charging time of EV_i. It is to be noted that the power consumed, P_{EV_i}, for all EVs of a particular value of battery capacity $E_{\max_i}$ is the same. As SOC_i for an EV_i decreases, the time CT_i required to charge it increases, and vice versa.

Next, the total EV load at a given time t is obtained. First, the EV charging pattern is obtained by the percentage distribution of EVs being charged throughout the day (found from data or surveys). Then the percentage of the EVs, $d\%$, being charged at time t, is multiplied by N, which is the total number of EVs of a particular type of vehicle throughout the test zone. This is then multiplied by P_{EV_i} to find the total EV load P_d at a given time t as shown in Eq. (2).

$$\begin{aligned} P_d &= P_{\text{EV}_i} \times N \times d\% \\ &= (1 - \text{SOC}_i) \frac{E_{\max_i}}{\text{CT}_i} \times N \times d\% \end{aligned} \tag{2}$$

3.2 Generating the Non-Electric Vehicle Load Curve

The next step is to obtain the non-electric vehicle load curve. The total load P_{tot} curve is first obtained, and the calculated EV load is subtracted from it to obtain the non-electric vehicle load curve, shown in Eq. (3). The non-EV load at a time instant t is denoted by P_n.

$$P_n = P_{\text{tot}} - P_d \tag{3}$$

3.3 Problem Formulation

Next, the load limit of the EV loads, P_{limit}, is calculated, to which the total EV load at a time t has to be confined to. This is found by subtracting the maximum non-EV load from the maximum power generation capacity of the power grid P_s. This is shown in Eq. (4).

$$P_{\text{limit}} = P_s - \max(P_n) \tag{4}$$

The determining function for the objective function of this problem is given by Eq. (5), which represents the power difference of the P_d from the load limit P_{limit}.

$$\begin{aligned} F &= P_{\text{limit}} - P_d \\ &= P_{\text{limit}} - \sum_{i=1}^{N} ((1 - \text{SOC}_i) \frac{E_{\max_i}}{\text{CT}_i}). \end{aligned} \tag{5}$$

In this paper, there are two objective functions. One is to increase the valley when the non-EV load is low (valley hour) and the objective function for this case is given by Eq. 6,

$$F_{\min} = \text{minimize } F \tag{6}$$

The second one is to reduce the peak when the non-EV load is high (peak hour) and the objective function for this case is given by Eq. 7,

$$F_{\max} = \text{maximize } F \tag{7}$$

Here, the constraints are

$$\begin{aligned}
&\text{subject to} \\
&F = P_{\text{limit}} - \sum_{i=1}^{N}((1 - \text{SOC}_i)\frac{E_{\max_i}}{\text{CT}_i}) \\
&P_d < P_s, \\
&0 < \text{SOC}_i < 1,
\end{aligned} \tag{8}$$

The constraint ($0 < \text{SOC}_i < 1$) sets the characteristic limits of the state of charge of an EV. However, as this is a completely random variable, this is not controllable and is not part of the optimization function. Constraint ($P_d > 0$) means that $(P_d)_{\text{opt}}$ is never zero.

The problem formulated with the objective functions and constraints is then optimized using the modified PSO to obtain the optimized EV load curve, $(P_d)_{\text{opt}}$.

3.4 Description of the Modified Particle Swarm Optimization (PSO)

The next step of the CCM model is the optimization of the non-coordinated EV load, which is carried out using the modified Particle Swarm Optimization (PSO) algorithm. The flowchart for this modified PSO algorithm is shown in Fig. 2.

For this model, the modification is introduced in the position updating phase of the optimization. The condition here is to optimize the EV load pattern according to the Non-EV load pattern. If the value of the non-EV load at time t is below the mean value of non-EV load pattern, the EV load at time t is maximized, but it is minimized if the vice versa is true. Within this operation, the state of the EV load is also taken into account. If the EV load at the non-EV peak hour is higher than the mean unoptimized EV load, then the EV load is lowered at that point, and the amount by which the position is reduced, D, is distributed equally among all EV loads that are below the mean EV load at all times t, meaning the valley hours of the EV load curve, $valley(P_{\text{EV}})$.

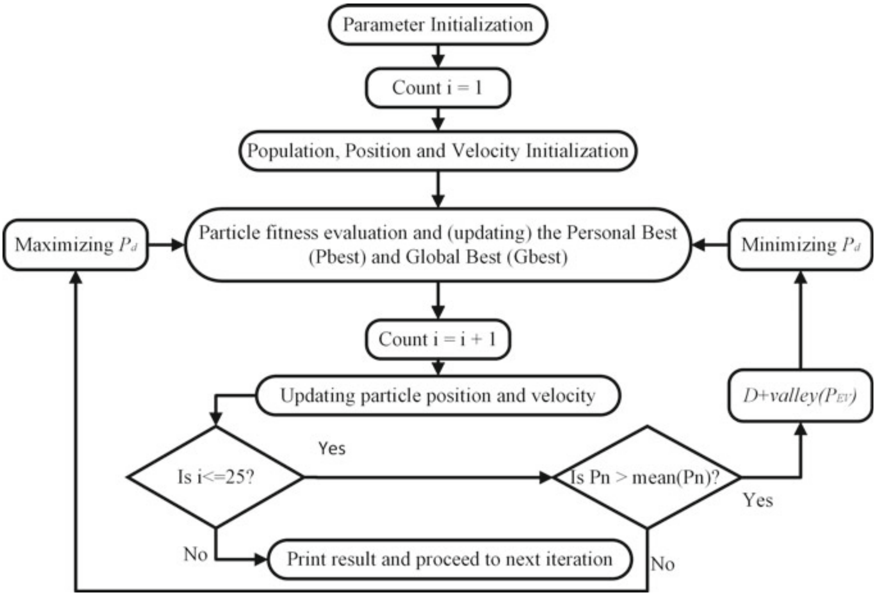


Fig. 2 Flowchart of the modified PSO algorithm

From this modified PSO algorithm, the optimized EV load curve is obtained, given by $(P_d)_{\text{opt}}$.

3.5 *Calculating the Charging Capacity Limit*

After $(P_d)_{\text{opt}}$ values have been obtained, the number of EVs that are allowed to be charged at different time intervals, N_{opt}, is calculated by dividing the $(P_d)_{\text{opt}}$ at those time intervals by the value of P_{EV_i}, shown by Eq. (9).

$$N_{\text{opt}} = \frac{(P_d)_{\text{opt}}}{P_{\text{EV}_i}} \tag{9}$$

Now, the total number of EVs being charged throughout the test zone for all times, t, is confined within the values of N_{opt}, which reduces any chance of the grid overloading. From here, the number of full capacity charging stations, C, required for each value of N_{opt} is calculated by dividing it by the number of EVs that a standard charging station can support, n.

$$C = \frac{N_{\text{opt}}}{n} \tag{10}$$

From the values of C, the maximum number of charging stations, $C_{\max}$, is determined, which is then considered to be the number of charging stations that should be constructed in the test zone. The N_{opt} is then divided by the $C_{\max}$ to find the number of vehicles that can be charged at each charging station at each time interval t.

$$N_c = \frac{N_{\text{opt}}}{C_{\max}} \tag{11}$$

With N_c values, the DSO can monitor the number of EVs in every charging station, to make sure that the limits of the model are being obeyed.

3.6 Mathematical Model with Respect to Easy-Bikes in Bangladesh

Coordinated charging of Easy-bikes using the proposed Charging Capacity Model (CCM) is expected to help alleviate many of the problems faced by all the stakeholders. These are the government, the DSO, the Easy-bike owners, and the common people. Implementation of the CCM would mean the government regulation would be easier, while the peak shaving and valley filling will make operations easier for the DSO. The Easy-bike (EB) owners can cut their losses while the ordinary people would be the biggest beneficiaries.

To formulate the solution for the study case of Bangladesh, the variables in the mathematical model are related with the data in this regard. This is summarized in Table 1.

In this paper, the proposed model is applied for Easy-bikes in Bangladesh. The power consumed by each of the EVs, P_{EV_i}, is an average of 0.4 kW [17], as the $E_{\max}$ value is about the same for all Easy-bikes using the same type of batteries.

The percentile distribution for the charging pattern of Easy-bikes is based on survey from several Easy-bike users and the employees in charge of several existing charging stations. This percentage distribution, $d\%$, is shown in Fig. 3.

Table 1 Summary of values of mathematical variables for Bangladesh

Qualities	Notation	Value
Total number of easy-bikes (present estimate)	N	1 M [3, 4]
Power consumed individual easy-bikes	P_{EV_i}	0.0004 MW [17]
Generation capacity	P_s	21,000 MW [18]
Maximum non-EV load	$\max(P_n)$	12,500 MW
Load limit of EV loads	P_{limit}	8500 MW
Maximum number of EBs in a charging station	n	12

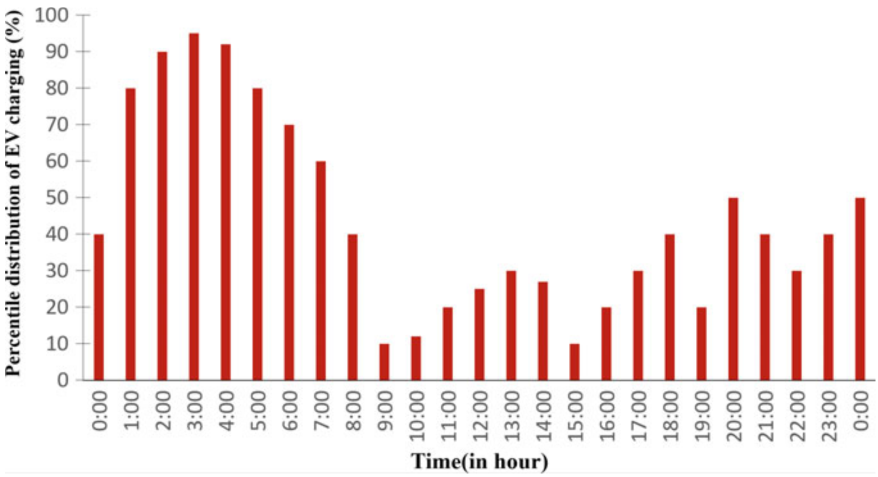


Fig. 3 Percentile distribution of easy-bike charging patterns in Bangladesh

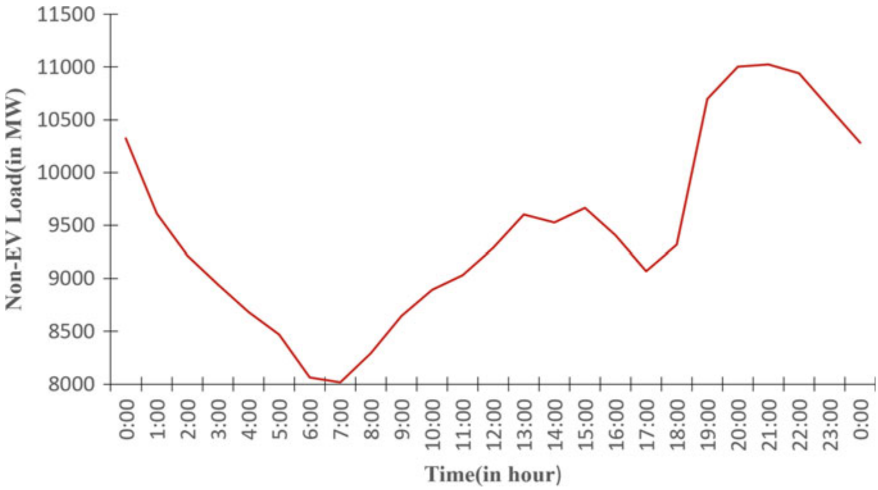


Fig. 4 Non-EV load curve

At present, total number of Easy-bikes present all over Bangladesh, N, is about 1 M [3, 4]. The value of P_d for Bangladesh for each time instant t for 24 h is then calculated using Eq. 2. The total load considered in this model is calculated by taking the average of the total load for 24 h for a typical day for every month, which are acquired from the official reports of the Power Grid Company of Bangladesh (PGCB) [19]. From here, the Non-EV load curve is found by subtracting the EV load curve from the total load curve, shown in Fig. 4.

Total generation capacity, P_s, of Bangladesh is currently at 21,000 MW [18]. The maximum non-EV load, $\max(P_n)$ is found to be about 12,500 MW in accordance with the load curve data used [19]. Thus for Bangladesh at present, P_{limit} has a value of 8500 MW. For Bangladesh, typical Easy-bike charging station can charge up to maximum of 12 Easy-bikes at a time. This information was acquired through survey, by contacting the owners of five known charging stations, in Gazipur, Jhenaidah, and Dhaka, over phone, keeping the COVID-19 situation in consideration. So n has a value of 12 in this case. It is assumed that this would remain the same for the new charging stations as well.

With these parameters set, the optimization operation is carried out on the data of Bangladesh, and the results are shown in the following section.

3.7 Communications Strategies

Communications are an important component in the CCM. This is segmented into two parts: the communications of the charging stations to and from the DSO, and the communications of the electric vehicles with the DSO. The first part is vital, as the DSO has to issue the number of electric vehicles allowed at each hour to all the charging stations consistently, as well as collect data from the charging stations. For this, Internet of Things (IoT) can be used to connect the charging stations with the DSO. This can be server-workstation network, with the central server being in the DSO. The charging stations are also connected to each other using IoT, in a parallel connection. The electric vehicles would be connected to the DSO through wireless communication through other IoT devices, and can view the vacancy status of their nearby charging stations based on the data provided by the DSO.

4 Result Analysis

When all the steps in the previous section are carried out with the data for Easy-Bikes in Bangladesh, the results for the optimized charging capacities are acquired. The results, of the charging power capacities and the charging station capacities, are described below.

4.1 The Results for Charging Power Capacities

Figure 5a shows the P_d load curve, as well as the $(P_d)_{\text{opt}}$ load curve, which has been generated using the modified PSO algorithm. It can be seen clearly that load shifting, consisting of both peak shaving and valley filling has taken place. One important reason why the algorithm did not lower the peaks to very low nor raise the valley too high, is that simply reversing the peaks and valleys to suit the needs of the power grid cannot be implemented practically. It is not feasible to completely lower

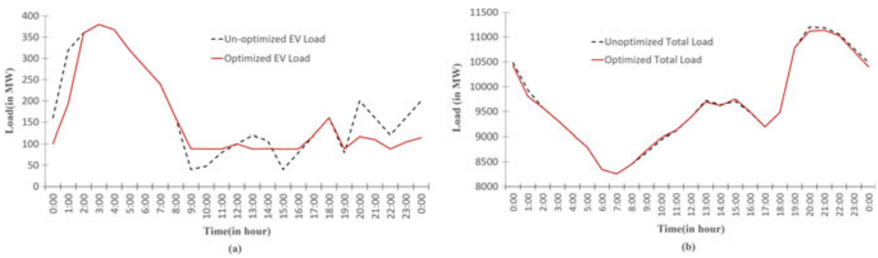


Fig. 5 Comparison of unoptimized and optimized load. **a** EV load, and **b** total load

the Easy-bike charging during the peak values at night, nor is it practical to force the EVs to be charged during the day, when most Easy-bikes are usually active. The goal of this optimization is to simply lower the number of EVs trying to charge at the same time to prevent an overloading condition. The total load curves before and after optimization is shown in Fig. 5b.

It can be seen that the optimization of the EV loads has significantly lowered the peaks of the total load curve as well as raised the valleys. It can be noted that this change appears graphically small. However, it should be remembered that these results are in accordance with the present penetration level of Easy-bikes in Bangladesh. At present, the maximum EV load at any point is less than 450 MW, while the total load at any point of time throughout the day is a maximum of 12,500 MW in accordance with the data used for this optimization. For higher penetration levels and more advanced forms of EVs such as full electric cars and buses, which have much higher values of $E_{\max}$, the contribution from this model would be much more significant. Table 2 summarizes the results of the optimization algorithm for peak shaving and valley filling.

4.2 *The Results for Charging Station Capacities*

After the number of EVs allowed at each time instant t is obtained, it is necessary to find the number of charging stations needed to support these infrastructures. These charging stations, of course, have to be legitimate and under the regulation of the DSO, as well as being state of the art, capable of supporting communications and integration of Internet of Things (IoT) devices. By that reasoning, the total number of charging stations with maximum filled capacity required for all times t is found, in accordance with Eq. 9, as described in Sect. 3. From all the values of C, the maximum value, $C_{\max}$, is obtained, which is found to be 66,667, which indicates the required number of charging stations throughout the country.

The optimized highest number of EVs allowed throughout the country for different time instants, is shown in Fig. 6a, and the number of EVs allowed per charging station, is shown in Fig. 6b.

Table 2 Values of the parameters and capacities throughout the model

Criteria	Parameters and results				
Parameters	N	P_{EV_i}	P_s	P_{limit}	n
	1 M	0.0004 MW	21,000 MW	8500 MW	12
Results	$\min(P_d)$	$\max(P_d)$	$\min((P_d)_{\mathrm{opt}})$	$\max((P_d)_{\mathrm{opt}})$	$C_{\max}$
	40 MW	380 MW	88.1 MW	380 MW	79,167

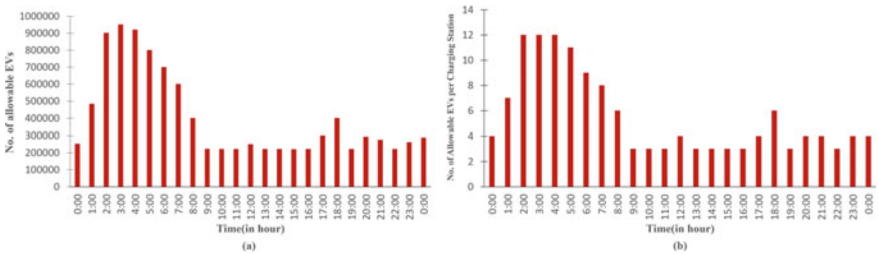


Fig. 6 Number of allowable EVs. **a** For all over the country at each hour throughout the day, and **b** for each charging station at each hour throughout the day

Table 2 summarizes the findings of the paper in terms on the charging station information.

4.3 Result Discussion

The results shown by this section clearly illustrates that the CCM satisfies both the objectives of this paper. For the case of Bangladesh, the optimized electric vehicle load curve are obtained for which peak shaving and valley filling are achieved, as shown by Fig. 5. The number of charging stations and the number of EVs allowed per charging station at every hour are also obtained, as shown in Fig. 6. This satisfies the second objective of the paper. The results are summarized in Table 2.

5 Conclusion and Future Direction

Most countries in the world are beginning to adopt EVs in a large scale, and as EVs are expected to completely replace the ICE-based vehicles in the not so distant future, the overloading of the grid due to mass charging of EVs is unavoidable. The CCM, as proposed by this paper, is well suited for test areas where large numbers of EVs have to be charged. This model is demonstrated for the Easy-bike charging in Bangladesh. The results, for the data of eleven months of 2021, show that this model successfully achieves peak shaving and valley filling for the EV load curve, thus reducing the peak of the total load, reducing the chances of overloading and load shedding. The adoption of this model is expected to ensure the grid stability while allowing higher EV penetration for similar test cases all over the world, as it is relatively easy to implement and very much suited to practical needs of countries like Bangladesh.

This model does not have any provisions yet for the accepting and rejection of vehicles, for the individual charging stations, which will arise out of potential delay in the arrival of EVs, as well as the time of charging of each vehicle. Practically, an algorithm has to be in place for managing the individual EVs at the charging stations. Other types of EVs will also have to be considered for the practical application of this model. As well as this, the locations of the charging stations are also an important parameter that has to be calculated. The authors plan to work on these topics in future works.

References

1. Global EV outlook 2021. Technical report, International Energy Agency (2021)
2. Is Bangladesh ready for the adoption of electric vehicles? LightCastle Partners. https://www.lightcastlebd.com/insights/2021/06/the-current-shape-of-ev-and-hybrid-vehicle-ecosystem-in-bangladesh. Last accessed 11 Oct 2021
3. Plying of easy bikes on highways needs to be controlled: Tawfiq-e-elahi. https://www.tbsnews.net/bangladesh/transport/plying-easy-bikes-highways-needs-be-controlled-tawfiq-e-elahi-214963. Last accessed 14 Oct 2021
4. Awal, M.R., Rahman, M.M., Islam, A.K.M.N., Al-Hossain, J., Khan, M.Z.R.: Energy consumption analysis of electric three wheelers in Bangladesh. In: 2019 IEEE International Conference on Power, Electrical, and Electronics and Industrial Applications (PEEIACON). IEEE, Dhaka, Bangladesh (2019)
5. 1m 'easy-bike' in country. https://thedailynewnation.com/news/149439/1m-easy-bikes-in-country.html. Last accessed 11 Oct 2021
6. Bangladesh plans to attract EV manufacturers. PV Magazine. https://www.pv-magazine.com/2020/12/04/bangladesh-plans-to-attract-ev-manufacturers/. Last accessed 13 Oct 2021
7. PDB for bringing battery-run vehicles under tariff regulations. https://www.dhakatribune.com/bangladesh/power-energy/2017/09/27/pdb-bringing-battery-run-vehicles-tariff-regulations. Last accessed 13 Oct 2021
8. Easy bike: a blessing or a curse. https://www.newagebd.net/article/58639/easy-bike-a-blessing-or-a-curse. Last accessed 13 Oct 2021
9. The Project on the Revision and Updating of the Strategic Transport Plan for Dhaka. Dhaka Transport Coordination Authority (DTCA) (2015)
10. BEPRC moves to formulate guideline for easy bikes. https://www.thefinancialexpress.com.bd/national/beprc-moves-to-formulate-guideline-for-easy-bikes-1615472819. Last accessed 12 Oct 2021
11. Quader: easy bike, 3-wheelers to come under purview of registration. https://www.dhakatribune.com/bangladesh/2021/02/18/quader-easy-bike-3-wheelers-to-come-under-purview-of-registration. Last accessed 13 Oct 2021
12. Clement, K., Haesen, E., Driesen, J.: Coordinated charging of multiple plug-in hybrid electric vehicles in residential distribution grids. In: 2009 IEEE/PES Power Systems Conference and Exposition. IEEE, Seattle, WA, USA (2009)

13. Masoum, A., Deilami, S., Moses, P., Masoum, M., Abu-Siada, A.: Smart load management of plug-in electric vehicles in distribution and residential networks with charging stations for peak shaving and loss minimisation considering voltage regulation. IET Gener. Transm. Distrib. **5**(8), 877 (2011)
14. Usman, M., Tareen, W.U.K., Amin, A., Ali, H., Bari, I., Sajid, M., Seyedmahmoudian, M., Stojcevski, A., Mahmood, A., Mekhilef, S.: A coordinated charging scheduling of electric vehicles considering optimal charging time for network power loss minimization. Energies **14**(17), 5336 (2021)
15. Yang, J., He, L., Fu, S.: An improved PSO-based charging strategy of electric vehicles in electrical distribution grid. Appl. Energ. **128**, 82–92 (2014)
16. Nguyen, V.L., Tran-Quoc, T., Bacha, S., Nguyen, B.: Charging strategies to minimize the peak load for an electric vehicle fleet. In: IECON 2014—40th Annual Conference of the IEEE Industrial Electronics Society. IEEE, Dallas, TX, USA (2014)
17. Zulkefa Rian, Md., Mizanur Rahman, A.N.M.: Study on power consumption and social aspects of battery operated auto-rickshaw. In: International Conference on Mechanical, Industrial and Energy Engineering 2014. KUET, Khulna, Bangladesh (2014)
18. Power generation in Bangladesh: important facts to look at. https://www.thedailystar.net/opinion/news/power-generation-bangladesh-important-facts-look-2052261. Last accessed 9 Oct 2021
19. Daily Operational Report. Power Grid Company of Bangladesh Limited. http://pgcb.gov.bd/site/page/0dd38e19-7c70-4582-95ba-078fccb609a8/-. Last accessed 15 Oct 2021

Designing of an Underwater-Internet of Things (U-IoT) for Marine Life Monitoring

Asif Sazzad, Nazifa Nawer, Maisha Mahbub Rimi, K. Habibul Kabir, and Khandaker Foysal Haque

Abstract Marine life and environmental monitoring of deep sea have become a major field of interest for quite a long time because of the immeasurable region of the area of the ocean that comes with its own dynamics and vulnerabilities. Creating the Underwater-Internet of Things (U-IoT) model within Underwater Wireless Sensor Network (UWSN) provides the scope of ensuring proper marine life monitoring which supports the aspects of 4th Industrial Revolution. The U-IoT network model is designed for an automated, efficient, smart process of data transfer for both underwater and overwater communications through acoustic waves and Radio Frequency (RF) data transfer techniques, respectively. The proposed U-IoT network model is created with an optimum number of autonomous underwater vehicles (AUVs) and surface sinks in order to address Bangladesh's overfishing problem (e.g., hilsa overfishing problem) which guarantees efficient management of the banning period by the authority. The network model is evaluated by comparing different deployment methods of AUVs and surface sinks taking the South Patch region of Bay of Bengal as the target area. The result shows that the proposed model transfers adequate data of marine life motion from the seafloor can enhance efficient administration of the overfishing problem.

A. Sazzad · N. Nawer · M. Mahbub Rimi · K. Habibul Kabir (✉)
Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Dhaka 1704, Bangladesh
e-mail: habib@iut-dhaka.edu

A. Sazzad
e-mail: asifsazzad@iut-dhaka.edu

N. Nawer
e-mail: nazifanawer@iut-dhaka.edu

M. Mahbub Rimi
e-mail: maishamahbub@iut-dhaka.edu

K. Foysal Haque
Northeastern University, Boston, MA 02115, USA
e-mail: haque.k@northeastern.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_21

Keywords Underwater-Internet of Things · Underwater Wireless Sensor Network · 4th Industrial Revolution · Marine life monitoring · Acoustic wave · AUV · Surface sink

1 Introduction

Marine life environment greatly impacts aquaculture, ecosystem which is source of biodiversity, food, life, etc., [1]. A healthy marine environment is critical for the fishing industry also a reliable source of animal protein for many people around the world whose income heavily depends upon the ocean accelerating influence in the global food economy. However, due to uncontrolled population growth, climate change, over-fishing, coastal pollution, eutrophication, etc., oceans are suffering, jeopardizing future food security. This necessitates marine life monitoring for a healthy ecosystem, as well as sustainable economic growth and food production.

Numerous studies on marine life monitoring are conducted using parameters, i.e., radioactivity monitoring [2], biodiversity measurement [3], underwater noise measurement [4], measurement of the effect of contaminant [5], coastal and oceanic observation [6], etc. In order to collect and deliver marine life data, an Underwater Wireless Sensor Network (UWSN) is essential to measure these parameters.

Therefore, in this paper, to build such an UWSN, Underwater-Internet of Things (U-IoT) is suitable and effective method. U-IoT enables continuous collection, send and aggregation of data on physical parameters of the marine life environment through underwater-IoT and transfer it for further processing through overwater-IoT (traditional Internet of Things). Deployment of U-IoT tremendously impacts the aspects of Fourth Industrial Revolution to create a smart network. It supports the analysis using machine learning and artificial intelligence in order to extract viable information for utilization in the Underwater Marine Life Monitoring system .

This paper proposes an efficient U-IoT network architecture to monitor marine life environment combining the motion detection of underwater objects with sensor nodes, autonomous underwater vehicles (AUVs), surface sinks and offshore sinks that optimizes data transfer time delay. The U-IoT network is modeled integrating both underwater and overwater communications. The data flow from sensor nodes to the surface sinks involves underwater acoustic communication, and standard terrestrial mode of communication, radio frequency, is used for data transfer from surface sinks to offshore sinks incorporates overwater radio frequency (RF) communication. Here, sensor nodes (deployed at the seafloor) generate data and transfer to next layer AUVs above them. The data is finally transferred to off-shore sinks through the surface sinks as shown in Fig. 1.

For underwater sensor network, acoustic waves are used as means of data transfer instead of traditional terrestrial methods of communication, such as radio frequency (RF). Under the water, RF highly attenuates the transfer of data (i.e., low data rate) [7], which reduces the transmission range rapidly [8], whereas acoustic waves have a low rate of absorption that enables long range data transmission with adequate data rate [9] and thus is a reliable data transfer medium for the proposed U-IoT model. In the proposed model, the South Patch region of Bay of Bengal is taken as the target region where Bangladesh's overfishing problem (e.g., hilsa overfishing problem) is addressed that guarantees efficient management of the banning period. In the result section, it is shown that data generated at the seabed due to motion of the marine life is efficiently detected with the adequate data rate at the off-shore sink nodes, that aids the detection of overfishing problem.

This architecture model of the proposed U-IoT is explained in detail in the sections beyond. The proposed paradigm is briefly compared to recent work on UWSN and U-IoT technology in Sect. 2. In Sect. 3, the U-IoT network model is described using a reference region. Further in Sects. 4 and 5, detailed simulation model is illustrated with appropriate reasoning. Lastly, the conclusion and the proposed model's future scopes are discussed in Sect. 6.

The structure of the paper is as follows. Section 2 describes the literature review, Sect. 3 describes the network model of proposed U-IoT. Sections 4 and 5 illustrates the detailed simulation model with appropriate reasoning. Lastly, Sect. 6 concludes discussing the proposed model's future scopes.

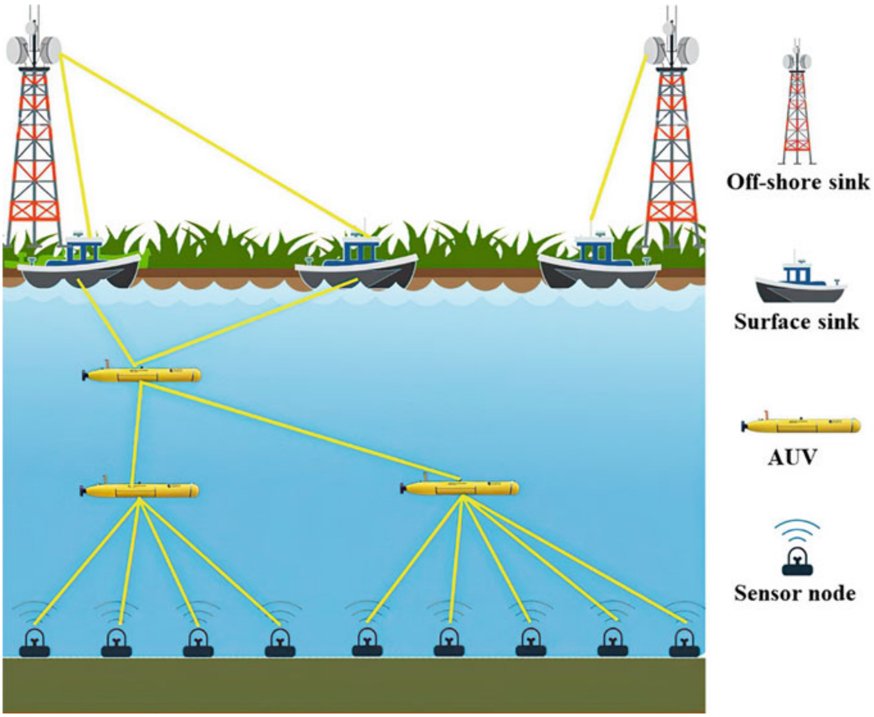


Fig. 1 Overview of U-IoT in monitoring marine life environment using acoustic wave for underwater and RF for overwater communication

2 Literature Review

Measuring different parameters (e.g., temperature, pressure, oxygen saturation, salinity, pH, Ammonia, etc.) of water to monitor marine environments with IoT-based smart technology [10] is challenging because most of the researches are done in a limited scale, i.e., small aquarium system or a fish pond, whereas the proposed approach allows for large-scale water quality monitoring.

Effective node deployment in the underwater seafloor is evaluated in a hierarchical UWSN using edge computing where Ant Colony-Based Efficient Topology Optimization [11] and Swarm optimization [12] are utilized to optimize deployment costs while maximizing network life and efficient node sensor range. An effective model for node battery life is proposed using Linear Regression Technique and Deep Neural Network modeling outperforming these techniques [13]. These articles mostly focus on finding node deployment positions with different optimization algorithms or modeling techniques optimizing node sensor energy, deployment cost maximizing network life, etc. The network we proposed is compatible with all the modeling and optimization algorithms.

In large-scale UWSNs, research on autonomous underwater vehicles (AUVs) is done with a cluster-based AUV-aided data gathering method to make a trade-off between energy savings and data transfer latency [14]. To solve this trade-off problem, a multi-hop realistic underwater sensor network model is simulated using acoustic wave [15] that suggests efficient AUV-aided underwater routing protocol (AURP) for improved data transmission reliability. The aforementioned studies focusing on AUV movement paths or clusterization can be collaborated with this research work suggesting the effective number of AUVs required.

Different parameters decide profitable routing algorithms. For UWSN, a number of energy-efficient and reliable MAC networks have been proposed [16]. A simulation for the U-IoT results in a high data packet transfer rate with low energy consumption and low latency [17]. In this study, an efficient model is simulated that proposes an architecture with the least amount of data transfer latency.

None of the studies cited above propose an optimal amount of surface sinks for reducing data transfer time. As a result, this paper is driven to create an underwater sensor network that proposes node cluster head orientation with the help of ocean vehicles and various number of surface sinks that maximizes data transfer time from node to off-shore sinks.

3 Network Model

The network model of U-IoT is obtained by deploying different underwater target devices such as nodes, AUVs, surface sinks and off-shore sinks. Firstly, the overall description of data transfer between these objects is explained. Later, a reference region is chosen in accordance with its importance of monitoring marine life with U-IoT.

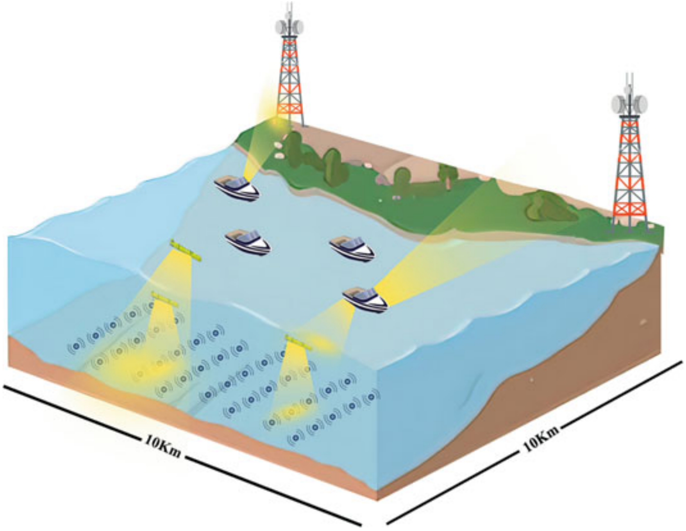


Fig. 2 3D view of U-IoT simulation model

3.1 *Description of UWSN Data Transfer Model for Marine Life Monitoring*

For marine life monitoring, the U-IoT is developed using nodes, AUVs, surface and offshore sinks as shown in Fig. 2. In the proposed model scenario, data is transferred from nodes to the AUV forwarded to the surface sinks that are in range; otherwise, the data is stored. Afterward, data is transferred from the surface sink to the off-shore sink, where it is collected in the nearest terrestrial network. As an example, Cox's bazar and Teknaf region of Bangladesh is used as reference to build the UWSN to create U-IoT.

3.2 *Proposed Scenario of Marine Life Monitoring: Case Study of Over-Fishing Problem of Hilsa in Bangladesh*

The preservation of Hilsa fishery resources is one of Bangladesh's most critical issues. Overfishing during the breeding season has put fishing stocks of Bangladesh in jeopardy. In order to address Bangladesh's overfishing problem, many new inventive ideas are developed by issuing a banning period where fishing is kept to halt. Despite government legislation, illicit and unauthorized fishing occurs during the ban period. However, existing techniques of controlling overfishing do not completely alleviate the problem because it is unable to monitor abnormalities during the dark hours. In order to be able to solve the predicament and provide an infallible solution, an

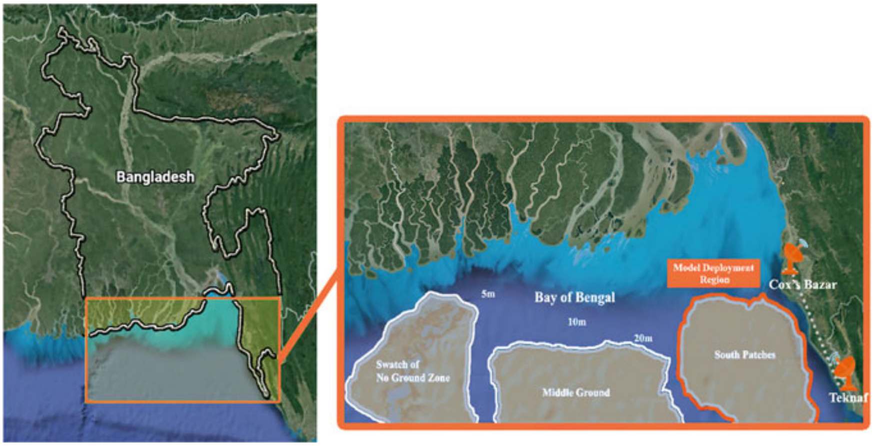


Fig. 3 Target region (Bay of Bengal-South Patch) for monitoring marine life

UWSN using U-IoT is built considering the South patch region of Bay of Bengal shown in Fig. 3

Bay of Bengal is taken into account as it is one of the favorable oceans with huge occurrence and variety of marine species with four fishing grounds which are (i) South patches, (ii) South of south patches, (iii) Middle ground, (iv) Swatch of no ground [18]. The UWSN for marine life monitoring system is conducted in the region of South patches which covers an area of 3662 km^2 and falls under Cox's Bazar region of Bangladesh [18] that can be utilized to monitor the underwater environment even at night, ensuring 24-h monitoring and control of the overfishing situation.

4 Simulation Model

The underwater sensors are positioned in a 10 km × 10 km region in the proposed model to make the calculation easier and a maximum depth of 100 m for the South patch region is taken into account which is shown in Fig. 2, where 46 nodes are placed on the seabed since the delivery rate and energy usage are both optimal [19]. The node sensor range is set at 35 m ensuring short-range acoustic communication [19]. Moreover, nodes are deployed in clusters where each cluster has a cluster head, and all other nodes in that cluster send their data to that cluster's cluster head, which communicates with AUV within its range and provides data with data generation rate of 1 kbps [11]. The nodes are placed at seabed both uniformly and randomly that will be further discussed in the next section.

To reduce packet loss, AUVs are deployed in the middle layer between the surface sink and nodes. Through graphical analysis, efficiency of AUV deployment is found that is further examined in Sect. 5.

Table 1 Parameters for marine life monitoring for Bangladesh's hilsa scenario

Oceanic parameters	Value
No. of nodes	46
Node sensor range	35 m
Data generation rate	1 kbps
AUV sensor range	1700 m
Depth of lower level AUV (from sea bed)	30 m
Depth of upper level AUV (from sea bed)	70 m
Speed of AUV	1.0 m/s
Boat sensor range	40 km (Sigfox)

For optimal data transfer to the surface sinks, the AUV sensor range employed in the simulation is set at 1700 m [20]. When the target items fall into the sensor range of each sender, they interact with one another. The upper-level target device within at a radius of 1700 m data is transferred from the lower level to the upper-level target device [20]. AUVs transport data to surface sinks and from surface sinks to off-shore sinks stationed in the study area in a similar manner. Also, it is assumed that the data collected from the AUVs will be sent from the boat through Sigfox. So, 40 km is considered as the sensor range for terrestrial communication from the surface sink to offshore sink [21].

Finally, two sinks are positioned at Teknaf and Cox's Bazar which will receive data from the surface sinks (considered boat in the proposed model) and will process it further.

The UWSN deploying U-IoT to monitor marine life is modeled using specific object parameters and their values which are given in Table 1. In the next section, the above-mentioned parameters are further explained using the Netlogo model environment in order to draw the appropriate conclusions from the simulation results.

5 Discussion of Simulation and Result Analysis

In this section, firstly the optimum node positioning at seabed is discussed along its graphical outcome. Secondly, the AUV deployment scenario involving layering of different number of AUVs is analyzed, and the result generating optimum use of the devices is concluded accordingly. In the similar manner, the number of surface sink usage is also carried out. Finally, using the optimal node, AUV and surface sink deployment a comparison analysis is given in the following subsections.

5.1 Uniform and Random Node Deployment in Seabed

Simulation is carried out using uniform and random node positioning in the U-IoT network model as shown in Fig. 4.

In case of uniform node deployment, greater rate of data transfer occurs at any time instance as a constant number of node is linked to the AUV for data transfer. On the contrary, in case of random node distribution, the data transfer rate is reduced as the number of nodes connected to AUVs at each time instance keeps changing. Sometimes, a few nodes are connected and sometimes none, causing this fluctuation of data rate transfer. Thus, from the simulation output as shown in Fig. 5, it can be deduced that data generated by the uniformly spaced node is received at the sink faster compared to when the network is randomly deployed.

As a result, the uniform distance within the node cluster head is taken into account because it depicts a closer relation to the ideal case scenario.

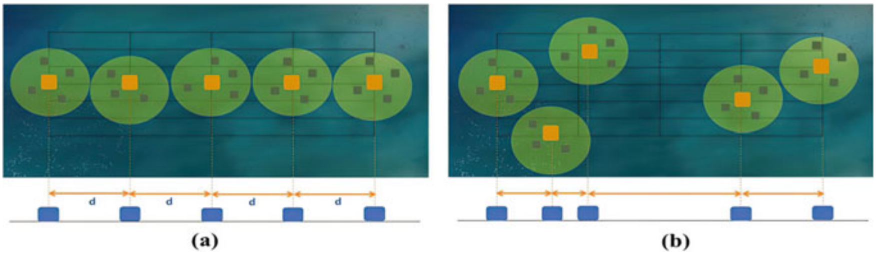


Fig. 4 Orientation of sensor cluster with **a** uniform cluster-heads, **b** random cluster-heads

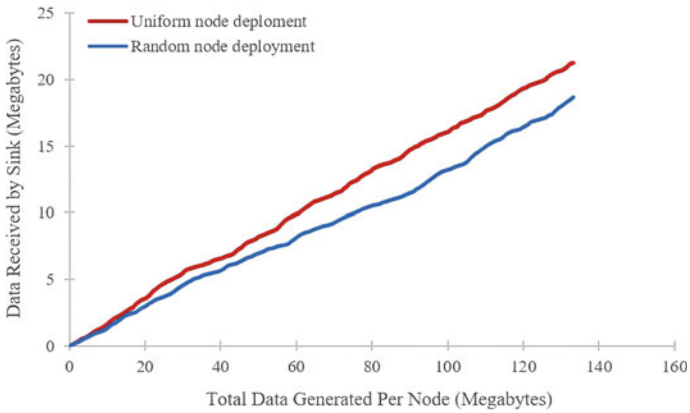


Fig. 5 Effect of uniform and random node deployment on network performance

5.2 *Optimal Deployment of Underwater AUV and Surface Sinks in UWSN*

In the middle layer, data transmission is done to the AUVs and from AUVs to the boats by implementing various scenarios as shown in Fig. 6.

In Fig. 7 analysis of different number of boat classifications used in both scenarios is shown, respectively.

The data sent by the AUV is limited to the maximum data packet of 5 kbps that can be transferred to the target devices causing data received by the upper level AUVs to get limited as well. Thereby, after installation of two boats, no matter how many surface sinks are introduced, the rate of data transfer between AUVs and boats does

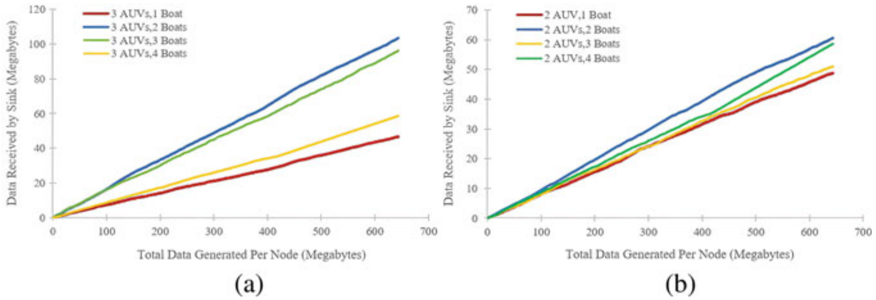


Fig. 6 AUV and surface sink localization in U-IoT with different number of boats, i.e., 1, 2, 3, 4. **a** Scenario 1: for 3 AUVs, **b** scenario 2: for 2 AUVs

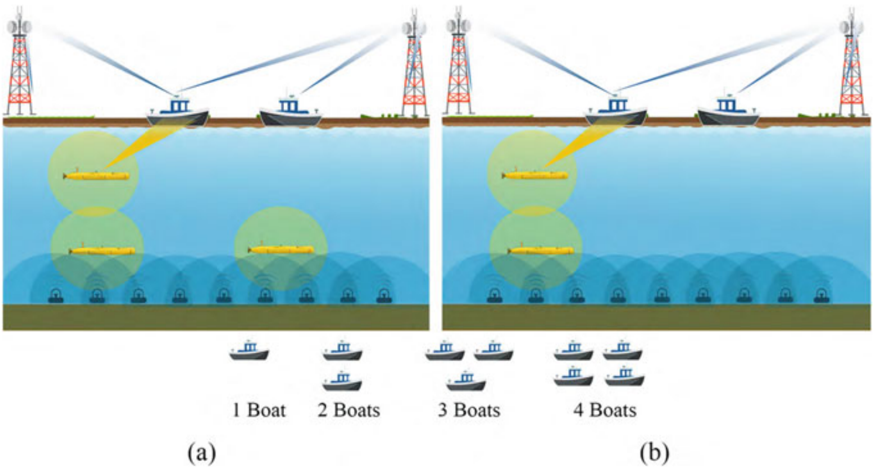


Fig. 7 Network performance. **a** Effect of scenario 1: for 3 AUVs, **b** effect of scenario 2: for 2 AUVs

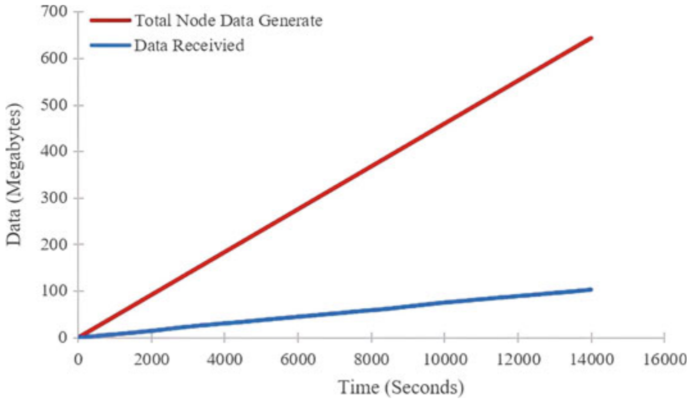


Fig. 8 Co-relation of data packet sent from the underwater sensor node and the received data volume at the overwater off-shore sink

not increase. Now, it can be seen from the simulation that using two boats results in faster transfer of data, regardless of the number of AUVs employed in two layers.

The main objective of this paper is to increase the rate of data transfer. So, when 3 AUVs are used, more data is received by the upper level AUV from the lower level AUV as an extra AUV is present in the bottom layer. Thus, from Fig. 7b, among the two scenarios, it can be concluded that the scenario with 3 AUVs gives the best flow of data to the surface sink.

Finally, the above-mentioned research proves that the scenario with three AUVs localized in two layers communicating with two boats provides the best possible data transmission in the U-IoT network model.

5.3 Data Transfer for UWSN Using U-IoT

From the analysis obtained above, a time-based comparison graph is plotted in Fig. 8 showing the total data generated by the nodes versus the total data received in the sink. The node data generation plot is linear as the data generation rate is taken to be constant as mentioned in Table 1. Due to the relative speeds of the mobile sensor devices (AUVs and surface sinks), link between two layers in the proposed UWSN deploying U-IoT is not formed at all times. And as data is only received when a link exists between two layers, a time delay is generated causing the received data graph to be a delayed plot.

5.4 Discussion of the Simulation Results for Marine Life Monitoring System

The optimal deployment of AUVs is carried out in two different scenarios, and the optimum number of surface sinks is found through creating four different scenarios in the above simulation analysis. Optimal deployment of AUVs and surface sinks in the proposed scenario refers to maximum data transfer rate from the underwater node sensors to overwater off-shore sinks.

The main objective is to reduce the overfishing problem in Bangladesh, which requires all day supervision through collection of marine life environment data. From the collected data, the population of adult hilsa fish can be monitored with the help of their motion detection. Moreover, if any kind of illegal activities take place, the sensor nodes will sense irregularity in underwater marine life motion detection. For example, if any fishing activities take place during the banning period, the motion of fishing net will be detected by the nodes providing variation in dataset. Thus, from the dataset record the extent of overfishing occurrence can be easily deduced to guarantee proper implementation of the banning process as the unauthorized fishing can be prevented through sufficient data transfer of marine life motion. Therefore, employing U-IoT within a short period of time allows accurate and faster data transfer ability, ensuring effective measure availability in the region whenever any irregularity is found in the received data.

6 Conclusion and Future Scopes

In this research, it is demonstrated that the UWSN architecture can implement the U-IoT network model efficiently, which can be used to monitor marine environmental conditions and collect specialized data. The results discussed above also provide proper evidence for the usage of optimal carrier deployment both underwater and overwater. This ensures an efficient, automated and reliable rate of data transfer through the U-IoT device positioning. Moreover, the U-IoT network model that is created in this paper also allows 24/7 surveillance of marine life. Thus, the overfishing problem in Bangladesh can be mitigated, ensuring proper regulation of the banning period through adequate transfer of data of marine life motion.

Furthermore, this architecture model ensures proper marine life monitoring in any part of the world, resulting in long-term ecosystem preservation. It is also evident that underwater sensor networking is an area with a lot of room for advancement in the future. Further, variables in IoT devices, like energy efficiency, storage capacity, memory installation, carrier band limitation and other data routing protocols can help to improve the model even more which will contribute to 4th Industrial Revolution.

Beyond the existing analysis of this paper, incorporation of efficient routing protocol to improve data delivery rate is the next focus of this research work.

References

1. Leape, J., William, E.P.: Harnessing the Fourth Industrial Revolution for Oceans. World Economic Forum (2017)
2. Byun, J.I., Choi, S.W., Song, M.H., Chang, B.U., Kim, Y.J., Yun, J.Y.: A large buoy-based radioactivity monitoring system for gamma-ray emitters in surface seawater. Appl. Radiat. Isot. **162**, 109172 (2020)
3. Muller-Karger, F.E., Miloslavich, P., Bax, N.J., Simmons, S., Costello, M.J., Sousa Pinto, I., Canonico, G., Turner, W., Gill, M., Montes, E., et al.: Advancing marine biological observations and data requirements of the complementary essential ocean variables (EOVs) and essential biodiversity variables (EBVs) frameworks. Frontiers Mar. Sci. **5**, 211 (2018)
4. Meyer, V., Audoly, C.: A parametric study of the environment and the array configuration for underwater noise measurement from ships in shallow water. In: Proceedings of the 26th International Congress on Sound and Vibrations (2019)
5. Williams, M., Kookana, R.S., Mehta, A., Yadav, S., Tailor, B., Maheshwari, B.: Emerging contaminants in a river receiving untreated wastewater from an Indian Urban Centre. Sci. Total Environ. **647**, 1256–1265 (2019)
6. Hall-Spencer, J.M., Harvey, B.P.: Ocean acidification impacts on coastal ecosystem services due to habitat degradation. Emerg. Topics Life Sci. **3**(2), 197–206 (2019)
7. Bernard, C., Bouvet, P.J., Pottier, A., Forjonel, P.: Multiple access acoustic communication in underwater mobile networks. In: 2021 Fifth Underwater Communications and Networking Conference (UComms) (2021)
8. Chen, J., Dai, Z., Chen, Z.: Development of radio-frequency sensor wake-up with unmanned aerial vehicles as an aerial gateway. Sensors **19**(5), 1047 (2019)
9. Huang, J., Wang, H., He, C., Zhang, Q., Jing, L.: Underwater acoustic communication and the general performance evaluation criteria. Frontiers Inform. Technol. Electron. Eng. **19**(8), 951–971 (2018)
10. Das, B., Ali, K., Memon, S., Shakoor, A., et al.: Monitoring of water quality of aquarium by using IoT technology. J. Appl. Eng. Technol. (JAET) **4**(2), 22–34 (2020)
11. Kim, S., Choi, J.W.: Optimal deployment of sensor nodes based on performance surface of underwater acoustic communication. Sensors **17**(10), 2389 (2017)
12. Duan, J.L., Lin, B., Cai, L.X., Liu, Y.X., Wu, Y.: Node deployment of marine monitoring networks: a multiobjective optimization scheme. Sensors **20**(16), 4480 (2020)
13. Reddy, T., Swarna Priya, R.M., Parimala, M., Chowdhary, C.L., Hakak, S., Khan, W.Z., et al.: A deep neural networks based model for uninterrupted marine environment monitoring. Comput. Commun. **157**, 64–75 (2020)
14. Ghoreyshi, S.M., Shahrabi, A., Boutaleb, T.: An efficient AUV-aided data collection in underwater sensor networks. In: 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA), pp. 281–288. IEEE (2018)
15. Yoon, S., Azad, A.K., Oh, H., Kim, S.: AURP: An AUV-aided underwater routing protocol for underwater acoustic sensor networks. Sensors **12**(2), 1827–1845 (2012)
16. Zenia, N.Z., Aseeri, M., Ahmed, M.R., Chowdhury, Z.I., Kaiser, M.S.: Energy-efficiency and reliability in MAC and routing protocols for underwater wireless sensor network: a survey. J. Netw. Comput. Appl. **71**, 72–85 (2016)
17. Arul, R., Alroobaea, R., Mechti, S., Rubaiee, S., Andejany, M., Tariq, U., Iftikhar, S.: Intelligent data analytics in energy optimization for the internet of underwater things. Soft Comput. **25**(18), 12507–12519 (2021)
18. Miah, M.S.: Climatic and anthropogenic factors changing spawning pattern and production zone of Hilsa fishery in the bay of Bengal. Weather Clim. Extremes **7**, 109–115 (2015)
19. Kim, S., Choi, J.W.: Optimal deployment of vector sensor nodes in underwater acoustic sensor networks. Sensors **19**(13), 2885 (2019)

20. González-García, J., Gómez-Espinosa, A., Cuan-Urquizo, E., García-Valdovinos, L.G., Salgado-Jiménez, T., Cabello, J.A.E.: Autonomous underwater vehicles: localization, navigation, and communication for collaborative missions. Appl. Sci. **10**(4), 1256 (2020)
21. Mekki, K., Bajic, E., Chaxel, F., Meyer, F.: A comparative study of LPWAN technologies for large-scale IoT deployment. ICT Exp. **5**(1), 1–7 (2019)

Beam Splitter Based on Dual-Core Photonic Crystal Fiber for Polarized Light Using Gold Metal Nanowires

Md. Ruhul Amin, S. M. Shahriar Nashir, Nazrul Islam, Md. Ahsan Habib, Mohammad Rubbyat Akram, and Sujay Saha

Abstract A beam splitter focused on a dual-core photonic crystal fiber with a hexagonal lattice structure has been designed. The Finite Element Method (FEM) is used to create and characterize the polarization splitter. The splitter can split polarized light into two polarization states at the wavelength of 1.55 μm. This work presents a working wavelength of about 1300–1950 nm. An extinction ratio that values lower than 10 dB with a total length of 850 nm can ensure an ultra-broadband bandwidth of 650 nm covering all communication bands. This proposed splitter may be capable of playing a vital role in communications and sensing applications.

Keywords Splitter · PCF · Coupling length · Extinction ratio · Fabrication

1 Introduction

The rapid growth of population and technology advancement demand more high speed and extensive capacity communications system. The PCF has unique characteristics. Polarization splitters can split light into two states (x-polarized and y-polarized light). Hence, the signal distortion can be reduced. The splitters focused on gold metal-filled PCF with all elliptical air, holes were designed, in 2015, with better bandwidth (1420–1980 nm) and a shorter length of 254.6 μm [1].

Wang et al. designed a square lattice structure with two elliptical air holes in the core output that were given in a short length of 93.3 μm, but a fast bandwidth of

Md. R. Amin (✉) · S. M. Shahriar Nashir · N. Islam · Md. A. Habib
Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh
e-mail: hmruhul16.mbstu@gmail.com

M. Rubbyat Akram
Deployment and Rollout Technology Division, Robi Axiata Ltd., Dhaka, Bangladesh
e-mail: rubbyat.akram@robi.com.bd

S. Saha
Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka, Bangladesh
e-mail: sujaysaha2906@ieee.org

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_22

70 μm with an extinction ratio of −79.3 dB at 1.55 μm wavelength [2]. Dou et al. published a paper in 2018 that described an energy transmission channel between dual-cores in a square lattice with two pitches [3] and obtained a device length of 103 μm with a bandwidth of 177 μm. He et al. designed a dual-core PCF using tellurite glass to get a short span of 89.05 μm with a bandwidth of 150 nm [4]. As it has four elliptical air holes with two ellipticity factors of 1.62 and 3.4, the structure is comparatively complex. Thus, the fabrication will be difficult. In 2019, Xu et al. designed a polarized splitter with a new approach using liquid and Ti infiltrated PCF having a short splitter length of 83.9 μm [5]. However in the reference level at −10 dB, the splitter represents a very ultra-scarce bandwidth of 32.1 nm.

According to the study, it is challenging to simultaneously design a splitter with an ultra-short length, a wide bandwidth, and a high extinction ratio. Because of the complicated elements of design characteristics and unavoidable terms between bandwidth and distance, the extinction ratio of specific splitters is relatively high but low bandwidth and vice versa. The proposed model shows the overall best performance in our work and can be used in optical communications systems compared to others.

2 Proposed Model and Theory

The cross-sectional view of the proposed model polarization splitter is shown in Fig. 1. The background material is silica, whose dispersion relationship is computed by the Sellmeier equation [6]. This PML will absorb all the reflections and light confinement in the entire computational area. The model consists of three types of air holes. Two of them are circular (sizes are small and large), and another one is elliptical. Here, d_1 and d_2 define the smaller and larger air hole diameters, respectively. And also, d_a and d_b determine the major and minor axis of elliptical air holes, and $\eta = d_a/d_b$ defines the ellipticity. Λ_1 and Λ_2 denote the pitch between two large circular air holes and between one small and one large air hole.

To create a dual-core, remove two small air holes along the y-axis and fill up two small air holes with gold metal along the x-axis (blue air holes in Fig. 1). The gold-filled air hole diameters d_g are equal to large air hole diameter d_2. From the improved Drude–Lorentz model, the relative permittivity of gold is computed as [7]

$$\varepsilon_m(\omega) = \varepsilon_\infty - \frac{\omega_D^2}{\omega(\omega - j\gamma_D)} - \frac{\Delta\varepsilon\Omega_L^2}{(\omega^2 - \omega_L^2) - j\Gamma_L\omega} \tag{1}$$

where ε_m, ε_∞, $\Delta\varepsilon$, and ω are denoting the relative permittivity, the relatively high-frequency permittivity of metal, the weighting factor, and the incoming light's angular velocity, respectively.

One of most important parameters of the coupling components of dual-core PCFs is the coupling length. For measuring the coupling length, the performance of the suggested splitter may be examined by utilizing four modes which are shown in

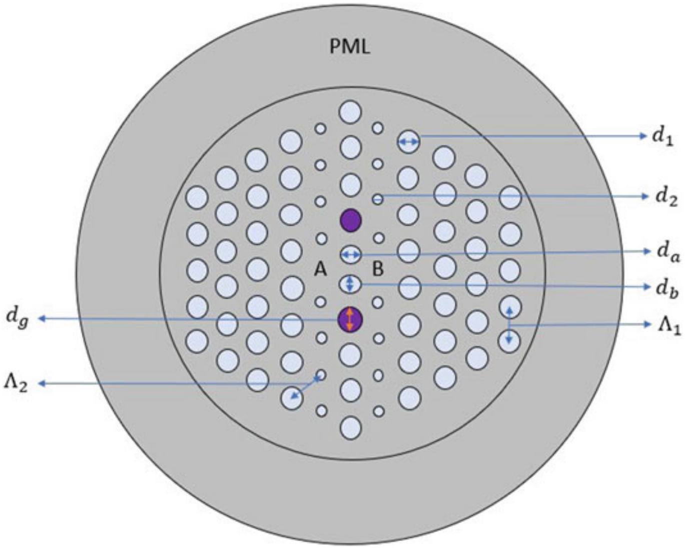


Fig. 1 Cross-sectiononal view of the splitter

Fig. 2, where Fig. 2 illustrates x-polarized and y-polarized light. Then, the following relation of coupling length can be expressed as [8]

$$L_{x,y} = \frac{\lambda}{2(n_{x,y}^{\text{even}} - n_{x,y}^{\text{odd}})} \tag{2}$$

here, L is the coupling length, and subscripts define the x- and y-polarization states. λ is the wavelength of the light. n indicates the effective mode index in the fundamental model, and the subscripts of n denote the polarization directions, and the superscripts indicate the light modes. For splitting the light into two cores, we have to maintain CL and CLR that meet the terms denoted as $L = mL_x = nL_y$, where m, n both are positive integers by propagating light from one core to another. The higher the coupling strength between the cores, the shorter the coupling length. The coupling length ratio that plays an vital role to achieve efficient beam splitter is defined by

$$\text{CLR} = \frac{L_y}{L_x} \tag{3}$$

CLR needs to be 2 (when $L_x < L_y$) or 1/2 (when $L_x > L_y$) to get optimum output within a short length from the splitter.

The light propagates from one core to another and for the input power in a specific core for particular polarized light, the output power can be calculated by following equation [9]:

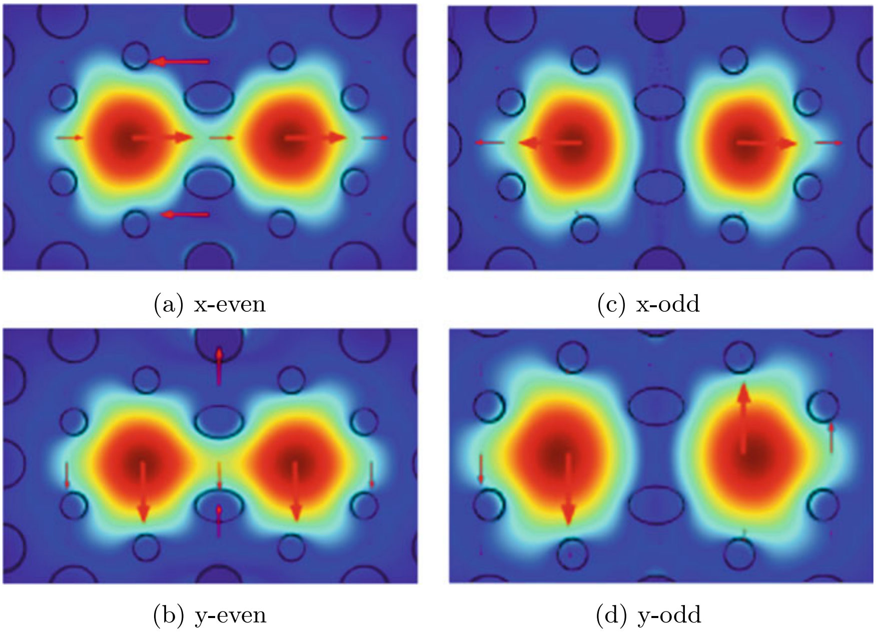


Fig. 2 Electric field for four polarization modes of the splitter

$$P_{out}^{x,y} = P_{in}^{x,y} \cos^2\left(\frac{\pi L}{2L_{x,y}}\right) \tag{4}$$

where P_{in} is the input power for a particular polarized light and L is the fiber length or length of the splitter.

In the polarization splitter, the key parameter is extinction ratio (ER), which can differentiate several types of beam splitter with different polarization states whether it is efficient or not. It is used to define working spectra of the splitter and can be expressed as [4]

$$\mathrm{ER} = 10\log_{10}\left(\frac{P_{out}^{y}}{P_{out}^{x}}\right) \tag{5}$$

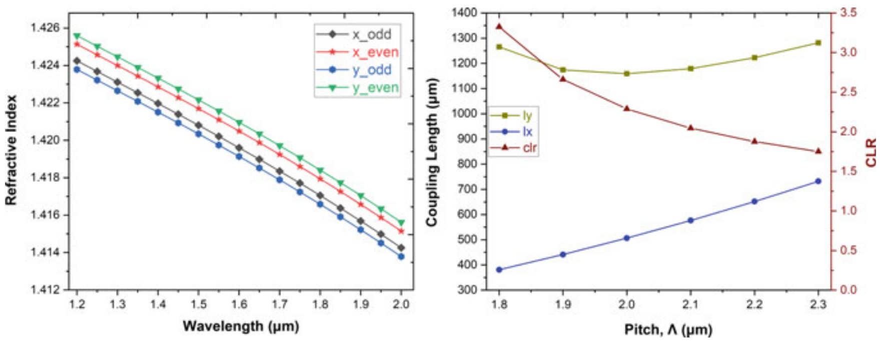
3 Optimization Process

The splitter's components are found using the COMSOL multiphysics simulation program. Figure 3a visually depicts the connection between effective refractive index and wavelength. The effects of using gold medals in the structure are described, corresponding to several figures.

For finding the value of CLR = 2 as well as getting better result, the diameters, pitch, ellipticity are optimized and analyzed at the wavelength 1.55 μm. The parameters of the mentioned PCF's initial values are as follows: Λ_1 = 2.1 μm, Λ_2 = 2.1 μm, $d_1 = d_g$ = 1.2 μm, d_2= 0.65 μm, d_a = 1.2 μm, d_b= 0.8 μm.

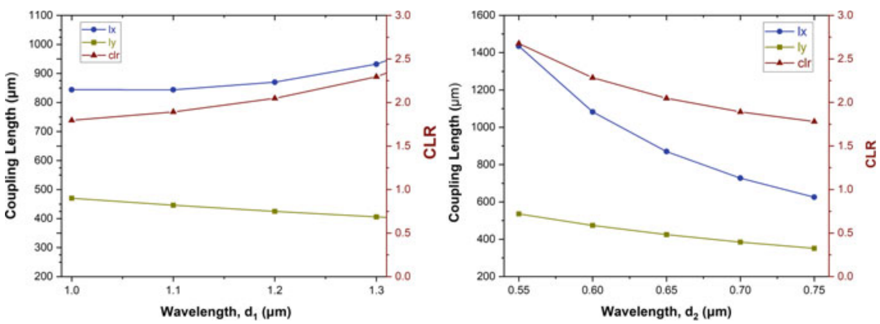
By changing Λ_1 and Λ_2 simultaneously, the effects of pitch variation are shown in Fig. 3b. In this situation, when Λ_1 and Λ_2 are altered from 1.8 to 2.3 μm by holding a fixed difference of 0.1 μm between Λ_1 and Λ_2 and other parameters will remain the same as before. By increasing pitch, the coupling length also increases, but the core power will be inconsistent. On the other hand, the pitch is more decreased, and the adjacent air holes may fall in a collision through the shorter length is always selected for splitters.After optimizing, CLR is 2.045 at pitch = 2.1 μm.

When the big air holes diameter (d_1) is varied from 1.0 to 1.4 μm, then the coupling length of x-polarized light(L_x) is increased from 844 to 1056 μm, and y-polarized light (L_y) is decreased from 470 to 387 μm which is shown in Fig. 4a as the diameter is increased.



(a) (b)

Fig. 3 Effective refractive indexes of core modes with respect to wavelength and electric field for four polarization modes of the splitter



(a) (b)

Fig. 4 Coupling length versus big and small air holes diameter

By decreasing the spaces between air holes bubbles, the bigger air holes improve energy transfer between the two cores, and so a larger diameter of d_1 shows a shorter coupling length. Therefore, the value of diameter (d_1) is optimized to 1.2 μm at CLR = 2.047.

When the diameter of the small holes (d_2) is decreased from 0.75 to 0.55 μm to enrich the birefringence, it is necessary to reduce pitch in each step of decreasing d_2 to maintain space between large and small air holes. Hence, the d_2 = 0.65 μm for achieving the optimized result which is shown in Fig. 4b.

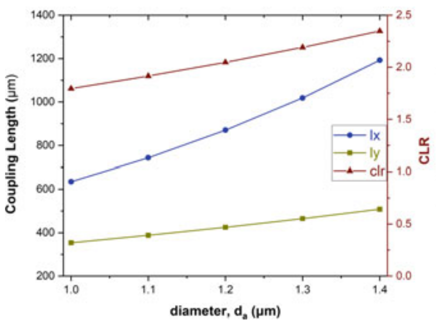
In Fig. 5a, changing the ellipticity (major axis) of the elliptical holes, which is varying from d_a from 1 to 1.4 μm when d_b is constant. Then, the coupling length increases, allowing more vital plasmonic interaction between the cores and the gold nanowires. However, the change in CLR also has to be considered while changing the ellipticity. When ellipticity is being changed, the CLR is 2.047 when da is equal to 1.2 m, which is incredibly near the required value of 2.

For Fig. 5b, changing the ellipticity d_b (minor axis) of the elliptical air holes is varying from 0.7 to 0.9 μm. The coupling length increases, which can allow more vital plasmonic interaction between the cores and gold nanowires. When ellipticity is being changed, the CLR is 2.047 when da is equal to 1.2 m, which is extremely near the required value of 2.

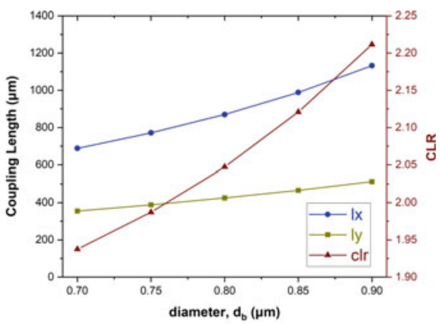
In other terms, the suggested design parameters' optimum values are Λ_1= 2.1 μm, Λ_2 = 2.1 μm, d_1 = d_g = 1.2 μm, d_2= 0.65 μm, d_a = 1.2 μm, and d_b = 0.8 μm for CLR = 2 at the wavelength of 1.55 μm.

The sensing applications' performance depends on the plasmonic material, so there must be careful about utilizing the material. There are several options to choose from among copper, silver, or gold as plasmonic material as these have higher resonance peaks. All of this gold provides a stable and optimal solution.

The effect of using gold metal in the proposed structure is shown in graphical format in Fig. 6a, where coupling length in left Y-axis and CLR in right Y-axis are increasing according to the wavelength increases, and the CLR is always nearly equal to 2. But in Fig. 6b, in the absence of gold metal in the proposed model PCF, the

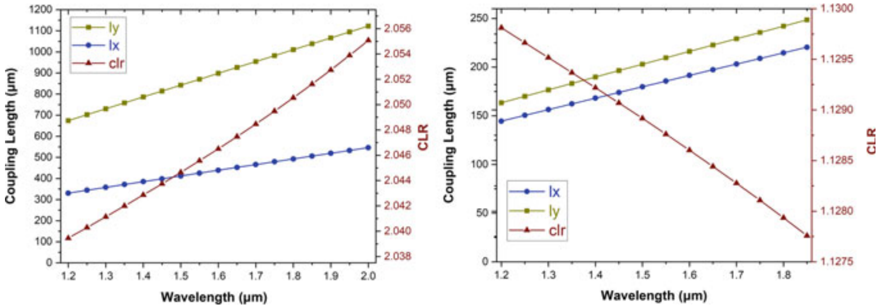


(a)



(b)

Fig. 5 Coupling length versus major and minor axis of the ellipse



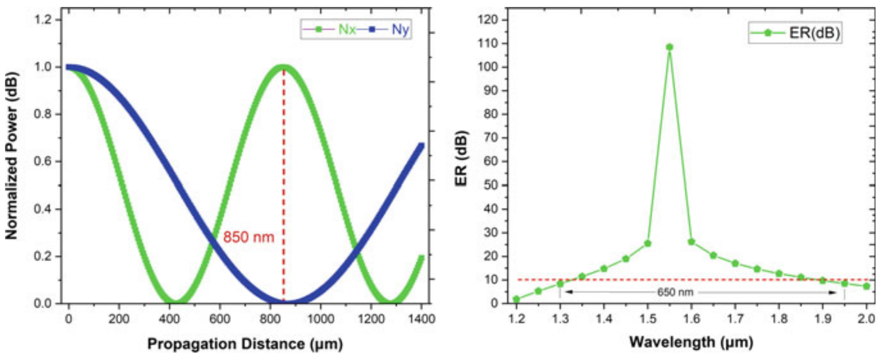
(a) (b)

Fig. 6 Coupling length and coupling length ratio (CLR) with respect to wavelength for the splitter

coupling length in left Y-axis increases, but CLR in right Y-axis can reach only the value of 1.13 instead of 2, which is not a desired optimal solution. Hence, the use of gold metal nanowires as plasmonic material significantly impacts transferring power between two cores. Otherwise, it isn't easy to complete the power transfer of any signal system or optical system.

4 Results and Discussion

If the normalized power in a core is maximum, the core will be minimal, and vice versa for dividing light into two cores. Because NP is a function corresponding to the propagation distance, the splitter length can be found through NP's output curve. Figure 7a shows that at a propagation distance of 855 μm, the NP of the x-polarized



(a) (b)

Fig. 7 Normalized power in core A presents the function of propagation distance and extinction ratio as a function of wavelength for optimized structure parameters ray is highest while the NP of the they-polarized glow is minimal. However, the polarized beams' power dwindles due to an ohmic drop in the gold metal nanowires.

The ER in the decibel unit is used to define the working spectrum with the ER more significant than 20 dB, shown in Fig. 7b. The splitter offers a very comparatively high ER of 108.52 dB at 1.55 μm and gradually falls with the wavelength, which can realize through Fig. 7b. The proposed splitter also tries to control the ER higher than 10 dB around the broad spectrum of about 650 nm from 1300 to 1950 nm, covering all essential communication bands for optimizing proposed structure parameters.

5 Fabrication

When the structure parameters are varying, then it is necessary to fabricate the tolerance of the structure. The parameters can vary up to ±2 and ±3%. However, the system's ability can be handled under ±1% [10] and the consecutive effects on ER according to the wavelength are shown in Figs. 8, 9, 10, 11, and 12. From Fig. 8, as the pitch, Λ in X-axis is lessening by 1%, 2%, and 3%, it is visible from Fig. 8 that the bandwidth lifts to 410 nm, 600 nm, and then decreases to 500 nm, respectively. A little bit of information is visible when the pitch increases about 3%, and short bandwidth is found as angles raised by 1% more than ever. The leading cause is that the CLR doesn't reach the desired value CLR, equal to 2.

In the proposed structure, the desired CLR value of 2 was found because of the combined effect of several optimized design parameters and using gold metal nanowires as plasmonic material. Furthermore, above the realization, it can be added that the optimized structure can control over ER around 10 dB at the wavelength of 1.55 μm for pitch fluctuation.

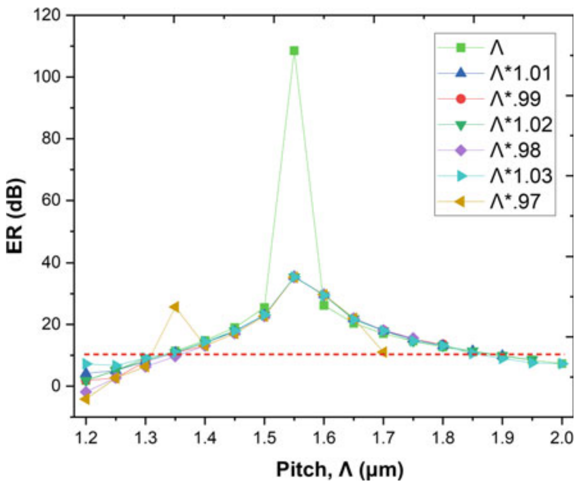
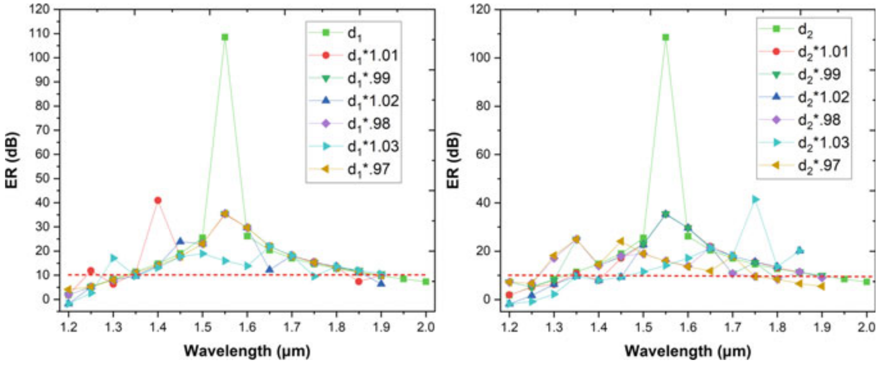
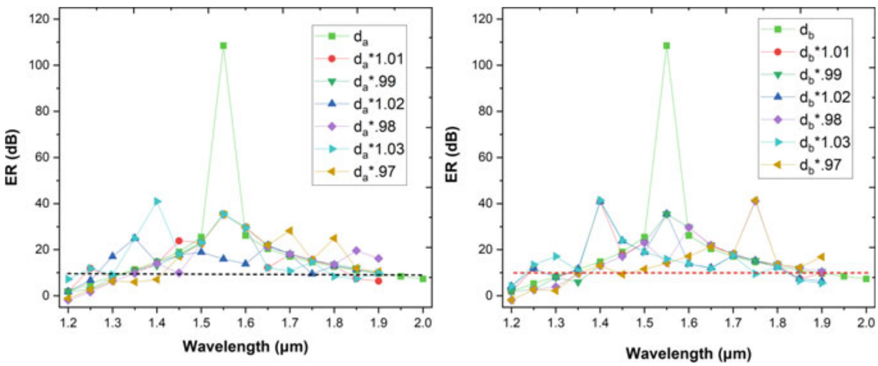


Fig. 8 Extinction ratio as a function of wavelength for the fabrication tolerance of pitch, Λ



(a) (b)

Fig. 9 Extinction ratio as a function of wavelength for the fabrication tolerance of large air hole diameter, d_1



(a) (b)

Fig. 10 Extinction ratio spectrum for various fabrication defects of major and minor axis, d_a & d_b

The ER spectra can be affected for the variation of the large air hole diameter d_1 by ±1, ±2, and ±3% are shown in Fig. 9a. The increment of d_1 by 1, 2, and 3% notably enhances the bandwidth, but the maximum ER shifts at the wavelength of 1.4, 1.2, and 1.6 μm separately. When d_1 is reduced by 1, 2, and 3% where bandwidth 200, 100, and 200 nm is generated, the splitter performance will also fall. This happens as the structure is changed with varying d_1. Thus, the splitter can't maintain the appropriate CLR of 2 under a bandwidth of 650 nm. The variation in the small hole diameters, d_2, can change the bandwidth of the splitter a few. This fabrication is shown in Fig. 9b. The effect of varying d_2 in the ER at 1.55 μm is that it is 35.46, 24.95, and 24.95 dB when d_2 is decreased by 1, 2, and 3% separately. Comparably, the ER of 35.27, 35.277 and 41.43 dB is generate because of the enhancement of d_2 by 1, 2 and 3% individually.

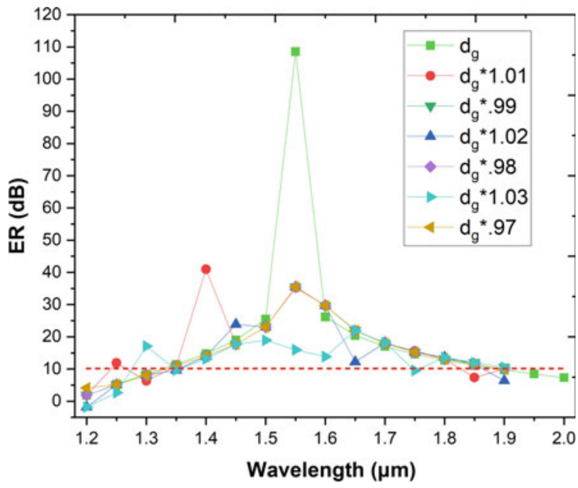


Fig. 11 Extinction ratio spectrum for fabrication tolerance of the diameter of gold nanowires, d_g

Changing the ellipticity of the major and minor axis creates almost the same splitter implementation as mentioned above, which is shown in Fig. 10a, b when the major and minor axis are reduced to 3%, then the maximum ER shifts to the wavelength of 1.45 µm and shifts to 1.5 µm when reduced by 2% and to 1.55 µm when declined by 1%. For increasing the major or minor axis about 1, 2, and 3% the peak ER is found at 1.4 µm wavelength at 10 dB reference level. So, the reduction of $d_a$or d_b by 2%, the ER is increased and the enhancement of $d_a$or d_b by 1, 2, and 3% individually. So, the above evaluation shows that the birefringence of the structure can be maintained through the varying in the ellipticity of elliptical air holes.

In Fig. 11, the fabrication tolerance of gold-coated air holes diameter slightly increases the bandwidth when gold-filled air holes diameter is increased by 2%. For the 2% decrement and 1%increment of d_g, the maximum ER will be shifted to 1.4 µm and the peak rest at 1.55 µm. Due to the change in gold-coated air holes diameter, the maximum ER will be shifted when d_g plays an ignorable effect on broadband of the splitter, which is shown in Fig. 11. Here, a small amount of gold nanowires shows an avoidable impact on the proposed model PCF. When all the parameters of the structure change simultaneously, then inspect mathematically for the utmost fluctuation of fabrication, as shown in Fig. 12. For the variation of Λ_1, Λ_2, d_1, d_2, d_a, d_b and d_g while changing these all parameters by ±1, ±2, and ±3% continuously at the same time, then ER curve will be shown in Fig. 12a, b. Among the three cases, an overall good result is generated while all structure parameters are increasing. But the change in all parameters can't reach the optimized desired output bandwidth. The bandwidth of almost 150 nm and for 1, and 2% incrementing range from the wavelength of 1.45–1.6 µm in Fig. 12b where this figure can show an optimal ER among all of this variation.

Discussing and realizing all of the fabrication tolerance impact on the splitter from Figs. 8, 9, 10, 11, and 12, it can be said that the splitter can handle it's own feature and mechanism with reasonable changes in structural parameters.

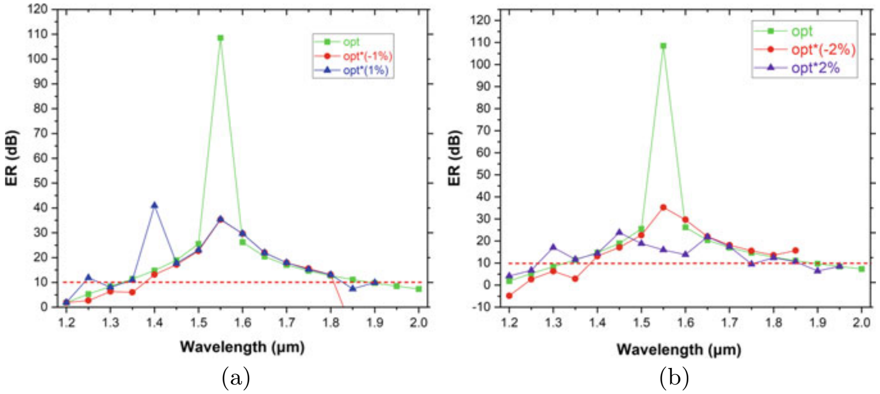


Fig. 12 Extinction ratio versus wavelength for the variation of all model parameters showing the resilience of the proposed splitter to fabrication tolerances

Table 1 Performance comparison between the proposed device and other works

Reference no.	Device length (μm)	ER (dB)	Bandwidth (ref. level)	Working spectrum (nm)
[1]	254.6	− 111	560 nm (< − 20 dB)	1420–1980
[2]	93.3	− 79.3	70 nm (− 20 dB)	–
[3]	103	− 73	177 nm (20 dB)	1508–1587
[15]	56.33	132.33	530 nm (20 dB)	1225–1755
[16]	481.3	− 98	200 nm (20 dB)	1450–1650
Proposed work	855	108.52	650 nm (< 10 dB)	1300–1950

The splitter process depends on the types of the beating of two orthogonal polarized photons' even and odd modes. Because the extension way of the coupling region and the splitter is uniform, the higher-level methods have not been realized. Stimulated in the suggested photonic crystal fiber structure splitter [11]. The existence of a higher-level mode in the suggested structure is also investigated: The higher-level mode (LP_{11}) has even and odd ways of x-, and y-polarized photons with lower practical index values and higher losses than the initial modes.

Due to the low device length and substantially increased efficiency, the suggested layer is added on a square lattice [12]. Furthermore, extremely birefringent fibers may be created by utilizing a square lattice with elliptical air holes of low ellipticity, which is better suited for single-mode transmission [13]. On the other hand, the working spectra of a PCF based on a square lattice are always higher than that of a PCF based on a triangular lattice [14]. As a result, PCFs on a square lattice can manage large amounts of data. More power, and it's easier to connect to other gadgets.

Finally, the suggested rending of splitters is likened. To that of existing splitters in the composition, it is summarized in Table 1. The size of various splitters is determined by splitter length, extinction ratio, bandwidth, and operating spectrum. Table 2 shows that there are just a few works [5, 8, 17]. Splitter bandwidth was calculated using a reference level of about 10 dB, while almost all used a 20 dB reference level. So the calculate bandwidth, we used a reference of 20 dB in our research. Some splitters are also noted for their short lengths and narrow bandwidths [2–5, 17] and some are broadband but large size [1, 8, 9, 18, 19]. Within the mentioned splitters, the suggested splitter has the lowest device length and the greatest extinction ratio. It also has a wide bandwidth that spans all communication frequencies, making it the most broad-spectrum splitter available. Our splitter model outperforms as originally stated splitters in terms of aggregate effectiveness.

6 Conclusion

Finally, this paper describes a dual-core PCF-based flatten and broadband polarization splitter. The FEM is used to optimize and quantitatively evaluate the structure. The suggested model's hexagonal structure is filled with gold for improved splitter performance based on the plasmonic resonance. The splitter is most effective at 1.55 μm wavelength, with a higher extinction ratio of 108.52 dB and a length of just 850 m. It can also sustain ER higher than 10 dB throughout a large spectrum of 650 nm from the wavelength of 1300 nm nm to 1950 nm nm, which covers many ranges of communication bands.

References

1. Khaleque, A., Hattori, H.T.: Ultra-broadband and compact polarization splitter based on gold filled dual-core photonic crystal fiber. J. Appl. Phys. **118**, 143101 (2015)
2. Wang, H., Yan, X., Li, S., An, G., Zhang, X.: Ultra-short polarization beam splitter based on dual core photonic crystal fiber. J. Mod. Opt. **64**, 445–450 (2017)
3. Dou, C., Jing, X., Li, S., Wu, J., Wang, Q.: A compact and low-loss polarization splitter based on dual-core photonic crystal fiber. Opt. Quantum Electron. **50**, 1–10 (2018)
4. Tao He, F., Juan Shi, W., Zhang, J., Hui, Z.Q., Zhan, F.: Polarization splitter based on dual-core photonic crystal fiber with tellurite glass. Optik **164**, 624–631 (2018)
5. Xu, Q., Luo, W., Li, K., Copner, N., Lin, S.: Design of polarization splitter via liquid and TI infiltrated photonic crystal fiber. Crystals (2019)
6. Tatian, B.: Fitting refractive-index data with the sellmeier dispersion formula. Appl. Opt. **23**(24), 4477 (1984)
7. Vial, A., Grimault, A.-S., Macías, D., Barchiesi, D., de la Chapelle, M.L.: Improved analytical fit of gold dispersion: application to the modeling of extinction spectra with a finite-difference time-domain method. Phys. Rev. B **71**, 085416 (2005)
8. Xu, Z., Li, X., Ling, W., Liu, P., Zhiyong, Z.: Design of short polarization splitter based on dual-core photonic crystal fiber with ultra-high extinction ratio. Opt. Commun. **354**, 314–320 (2015)
9. Khaleque, A., Mironov, E.G., Hattori, H.T.: Analysis of the properties of a dual-core plasmonic photonic crystal fiber polarization splitter. Appl. Phys. B **121**, 523–532 (2015)

10. Lo, C. W., Stefani, A., de Sterke, C. M., Blanco-Redondo, A.: Analysis and design of fibers for pure-quartic solitons (2018)
11. Lu, W., Lou, S., Wang, X.: Ultrabroadband polarization splitter based on a modified three-core photonic crystal fiber. Appl. Opt. **52**(35), 8494–500 (2013)
12. Rosa, L., Poli, F., Foroni, M., Cucinotta, A., Selleri, S.: Polarization splitter based on a square-lattice photonic-crystal fiber. Opt. Lett. **31**(4), 441–3 (2006)
13. Liu, Y., Lai, Y.: Optical birefringence and polarization dependent loss of square- and rectangular-lattice holey fibers with elliptical air holes: numerical analysis. Opt. Express **13**(1), 225–35 (2005)
14. Maji, P.S., Chaudhuri, P.: Near-elliptic core triangular-lattice and square-lattice PCFS: a comparison of birefringence, cut-off and GVD characteristics towards fiber device application. J. Opt. Soc. Korea **18**, 207–216 (2014)
15. Rahman, M.T., Khaleque, A.: Ultra-short polarization splitter based on a plasmonic dual-core photonic crystal fiber with an ultra-broad bandwidth. Appl. Opt. **58**(34), 9426–9433 (2019)
16. Shimu, S., Shawkat, M.T.B.: A polarization splitter using dual core photonic crystal fiber. In: 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021, pp. 1–6
17. Wang, J., Pei, L., Weng, S., Wu, L., Li, J., Ning, T.: Ultrashort polarization beam splitter based on liquid-filled dual-core photonic crystal fiber. Appl. Opt. **57**(14), 3847–3852 (2018)
18. Jiang, L., Zheng, Y., Hou, L., Zheng, K., Peng, J., Zhao, X.: An ultrabraoadband polarization splitter based on square-lattice dual-core photonic crystal fiber with a gold wire. Opt. Commun. **351**, 50–56 (2015)
19. Zhao, X., Hua, L., Xiong, Q., Jiang, G., Cheng, J.: Ultra-short and broadband polarization splitter based on PCF and metal surface plasmons resonance. Opt. Quantum Electron. **51**, 1–11 (2019)

Investigating the Performance of Delay-Tolerant Routing Protocols Using Trace-Based Mobility Models

Md. Khalid Mahbub Khan, Muhammad Sajjadur Rahim, and Abu Zafor Md. Touhidul Islam

Abstract The mobility patterns of nodes significantly influence the performance of delay-tolerant network (DTN) routing protocols. Trace-based mobility is a class representing such movement patterns of nodes in DTN. This research analyzes the performance of DTN routing techniques on trace-based mobility regarding delivery ratio, average latency, and overhead ratio. Three real traces: MIT Reality, INFOCOM, and Cambridge Imotes are implemented on five DTN routing techniques: Epidemic, Spray and Wait, PRoPHET, MaxProp, and RAPID. For more explicit realization, Shortest Path Map-Based Movement from synthetic mobility model has also experimented with the traces. The Opportunistic Network Environment (ONE) simulator is used to simulate the considered protocols with these mobility models. Finally, this research presents a realistic study regarding the performance analysis of these DTN routing techniques on trace-based mobility along with Shortest Path Map-Based Movement by considering the variation of message generation intervals, message Time-To-Live (TTL), and buffer size, respectively.

Keywords Delay-tolerant network · Routing protocol · Trace-based mobility model · Performance analysis · Opportunistic network environment simulator

1 Introduction

Delay-tolerant networking (DTN) architecture has been proposed to address the application drawbacks in dynamic networks, which may experience topological

Md. K. M. Khan (✉) · A. Z. Md. T. Islam
Department of Electrical and Electronic Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
e-mail: khalidmahbub.khan@yahoo.com

A. Z. Md. T. Islam
e-mail: touhid.eee@ru.ac.bd

M. S. Rahim
Department of Information and Communication Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
e-mail: sajid_ice@ru.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_23

diversity due to the intermittent connectivity among nodes. Several reasons exist behind such intermittent connections of nodes in DTN: bounded coverage range of nodes, extensive distribution of mobile nodes, restriction of resources (i.e., energy), higher interferences or other medium impairments, etc. These characteristics make the DTN scheme a perfect replacement for conventional Mobile Ad Hoc Network (MANET). MANET architecture is not suitable for those applications since an instantaneous path from source to destination nodes is required for transmitting data. Also, any delay handling mechanism to account for these delays because of intermittent connectivity of nodes is absent in MANET. Therefore, higher latency decreases the data delivery rate in these cases, whereas DTN architecture improves the success rate of data delivery by exploiting the "Store-Carry-and-Forward" strategy on mobile nodes suffering intermittent connections to each other. Within this approach, nodes store data on limited buffer storages, while connectivity is unavailable and bears data for an interim period to forward before the connection of nodes with the target node is available. Thus, DTN can be an excellent option to deliver data successfully in those networking scenarios intended to operate in heterogeneous mobile networks, interplanetary communications, rural communications, extreme and emergency terrestrial environments, viz. natural disaster-affected areas, etc. [1, 2].

Several routing protocols have been proposed for DTN scenarios to improve the data delivery rate considering minimum latency and overhead. The flooding-based protocol can be considered the first-generation protocol where the message is replicated uncontrollably. Epidemic [3] is an example of a flooding-based protocol. Later, message replication was restricted to next-generation protocols. Spray and Wait [4] is a notable instance of this type of protocol. Then, a probabilistic concept is adopted in the message routing strategy. Probabilistic Routing Protocol using the History of Encounters and Transitivity (PRoPHET) [5] protocol is an example of such probabilistic protocols. Resource allocation-based and schedule-based protocols are the posterior protocols. Resource Allocation Protocol for Intentional DTN (RAPID) [6] and MaxProp [7] are examples of such routing techniques. Although such protocol's application perspectives are different, researchers have been investigating better routing solutions for DTN scenarios.

However, the movement pattern of nodes' is an essential factor to consider in DTN architecture. The message transmission time depends on the nodes' movement pattern and the distance between the source and destination node. This transmission time may vary from a few minutes to more hours or multiple days. Accordingly, the performance of data delivery of the routing technique significantly depends on the selection of a proper mobility pattern [8]. Moreover, the frequency and duration of the contact are influenced by the movement pattern of nodes. The nodes' movement model depends on various factors such as the entire networking area, map of the street, and nodes' speed. Mobility patterns can be categorized into two types. The first type is synthetic mobility pattern which is a statistical model. In this case, mobility pattern is generated according to particular rules. Shortest Path Map-Based (SPMB) Movement [9] is one example of such a model. These synthetic mobility models are mainly specified for a group of users.

On the other hand, real trace-based mobility is another type of movement pattern. These real traces are developed from many real-life applications and exhibit the actual behavior of nodes in a specific environment. Several real traces are available in an online-based repository known as the CRAWDAD project [10]. Traces are established on contact traces and GPS traces. A finalized contact list is supported by a contact trace when two or more nodes are involved in a contact. Contact's start time and end time are also included there. Differently, the location of nodes is given for a specific duration in GPS traces. However, unfortunately, some application restrictions have been realized in real traces. New scenarios, which are outside the collections, are not supported by traces.

Furthermore, traces have a fixed size, so it is impossible to extend the size of the traces, and they are unable to support any variation from user contexts. So, traces would not consider if the users were moving in challenging and troubling environments. Apart from this, the utilization of traces in simulation is less cost-effective as external files are necessary to read the information related to contacts. In comparison, synthetic movement models offer additional functionalities. These models support different new scenarios and allow the parameters to be adjustable. Accordingly, if the change of any network-related attributes, viz., if the network is extended due to increasing node density, synthetic mobility models can afford that is absent in real trace-based mobility models [11].

This study has investigated a comparative performance analysis of several DTN routing protocols considering three real traces: MIT Reality, INFOCOM, and Cambridge Imotes. Shortest Path Map-Based Movement from synthetic mobility is also considered in our study for simplifying the comparison. Opportunistic Network Environment (ONE) Simulator [12] is used as the simulation tool to perform all the simulations. We have analyzed the performances of simulated protocols considering three traces and one synthetic mobility model based on three performance metrics: delivery ratio, average latency, and overhead ratio. The remaining portion of this paper consists of: Relevant literature study is provided in Sects. 2 and 3 which presents a brief description of considered mobility models in this research, and Sect. 4 gives a short review of experimented DTN routing protocols. An explanation regarding the simulation environment and necessary parameters is given in Sect. 5, whereas Sect. 6 analyzes the comparative performances of simulated routing protocols. Lastly, Sect. 7 contains the conclusions and some directions for future research works.

2 Related Works

Real trace-based mobility and synthetic mobility models are considered in many research works by researchers. Besides, performance analysis of DTN routing protocols from various perspectives also gained attention in various researches. In [13], the authors have considered performance metrics such as delivery ratio, latency, buffer occupancy, and hop count to evaluate the performance of several DTN routing protocols. Only Random Waypoint from synthetic mobility is used as the mobility model in this case. A comparison of several synthetic mobility models such as Shortest Path Map-Based, Map-Based Movement, and Random Waypoint is presented from various perspectives in [14]. Nevertheless, real trace-based mobility models are unaddressed here. Furthermore, the authors in [15] have demonstrated the energy consumption of the existing DTN protocols using three synthetic mobility models, namely Random Walk, Random Waypoint, and Shortest Path Map-Based movement. Performance comparison of those synthetic models for simulated DTN protocols is also given. Although several mobility models are examined, the absence of any real traces can be considered the limitation of this work. On the other hand, two experimental scenarios are considered to analyze the performance of four different conventional DTN routing techniques on the Tirana city map based on four performance metrics in [16]. Except for any real traces, Shortest Path Map-Based model is chosen here. In [17], a performance analysis of social-based routing protocols with traditional DTN routing protocols is investigated. This study considers only the Shortest Path Map-Based Movement model, which is an example of a synthetic mobility model. This performance analysis is conducted using three performance metrics: delivery ratio, average latency, and transmission cost. In addition, the impact of increasing node density and TTL is considered. However, the trace-based mobility models are excluded from this research. Another performance comparison of synthetic mobility models is exhibited in [18] without including any real trace. In this study, existing DTN routing techniques are involved. Moreover, performance analysis of several DTN routing protocols along with the energy efficiency is studied in [19]. Like the previous works, real traces were absent in this study, while only the Shortest Path Map-Based Movement model is adopted as the default mobility model. These pitfalls of the above works motivated us to investigate the impact of real traces on the performance of DTN routing approaches.

3 Mobility Models Under Investigations

In our study, we have included three real trace-based mobility models: MIT Reality, INFOCOM, and Cambridge Imotes. Shortest Path Map-Based Movement from synthetic mobility model is also considered in this research to get a comparative idea about the outcomes. A brief explanation of the above-mentioned mobility models is given in this section.

MIT Reality is a trace that is campus-oriented. It is based on 97 Nokia 6600 cellphones which are interconnected through the Bluetooth interface. In this trace, the contact information was obtained from the course-wise faculty members and MIT students in the 2004–2005 academic year. This contact information was formed not only from the data transmission among participants but also from the location and proximity of the participants. The total duration of this trace is 246 days [20].

INFOCOM trace is also recognized as Haggle 3 project trace. This trace originated from a conference having the same name and sponsored by IEEE. This conference was organized in 2005. Here, 41 volunteering participants carried the same number of Intel Motes (Imotes) devices through Bluetooth connectivity. Later, the organizers associated 98 participants with the same number of Imotes devices in 2006, and the dataset found from there is known as INFOCOM 6. The entire duration was 3 days for both INFOCOM 5 and INFOCOM 6. INFOCOM 5 is endorsed within this study so we would use INFOCOM to identify INFOCOM 5 trace [11].

Cambridge Imotes is another campus-based trace. In this trace, 36 Imotes were provided to the same number of participants in the University of Cambridge campus area. Such devices were set up in the most well known and trendy places on campus for 10 days. Bluetooth interface was used here as the communication technology. It is also familiar as Haggle 4 project trace [11].

Shortest Path Map-Based Movement is a synthetic mobility model. It is a more realistic model than other map-based movements. Here, map data that is fixed provides random paths where nodes are moving randomly. Random points on the map are chosen by the nodes based on their recent positions. Nodes further try to discover and pursue the shortest path to the random points. The current locations construct the shortest path of these random points following Dijkstra's algorithm. The target random points are randomly selected, or a list known as a Point of Interest (POI) is formed. This list is chosen depending on the popular real-life places, viz., well-known tourist places, shops or restaurants, etc., which are prominent and trendy [12].

4 Investigated DTN Routing Protocols

We have experimented with five DTN routing protocols: Epidemic, Spray and Wait, PRoPHET, MaxProp, and RAPID in this research. This section describes the basic mechanisms of these protocols.

In Epidemic protocol, uncontrolled replication of messages takes place. In this strategy, the source node replicates the message it has continuously. Then, it forwards the message copies to those encountered nodes that have not received any copy of this message yet. In such a manner, the intended target node will receive the message copy eventually. The main drawback of this routing strategy is that multiple copies of the same message will flood the network. Furthermore, another shortcoming of Epidemic is that the concept of limited resource utilization (i.e., bandwidth, buffer space, etc.) is absent in this approach. So, to deliver data successfully in a network, it will consume a significant amount of network resources. This deficiency makes Epidemic the worst candidate for routing in DTN [3].

Spray and Wait protocol exploits the limited replication concept in message forwarding. In this case, a limited number of message copies are forwarded to a similar number of relay nodes. Source node will forward L copy of the message to the *L* relays. Here, *L* denotes the maximum number of message replicas that are allowed to be dispatched. The value of L is influenced by some factors, viz. network's node density, etc. Spray and Wait strategy incorporates two phases [4]:

Spray Phase: In the "Spray" phase, a message-carrying node will forward L replicas of a message to the L number relay nodes.

Wait Phase: L relay nodes will wait by carrying L replicas of the message until their expected encounter with the destination node to deliver the carried replicas of the message. This phase is known as the "Wait" phase.

Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET) is a forwarding protocol where the message carrier node maintains a set of probabilities. Nodes' actual real-world confrontation likelihood is considered in this approach. While a meeting with a message carrier source node will take place, a message will be forwarded to the encountered nodes depending on probabilities. The encountered node which has a higher chance of delivery will obtain the message first [5].

In the MaxProp protocol, specific schedules are assigned and prioritized to forward data packets. A message carrier node maintains a ranked list of data packets to determine which packet will be forwarded or dropped on a primitive basis. This ranked list depends on the cost computed from the source node to all destined target nodes and reflects the message delivery probability estimation. The packet preference is maintained in the MaxProp protocol depending on the minimum hops count by avoiding the multiple version reception of the same message [7].

The resource allocation problem of the DTN environment is addressed in Resource Allocation Protocol for Intentional DTN (RAPID) protocol to improve the routing performance. In such a case, DTN architecture is assumed as a scheme that is utility-driven. This utility-driven scheme maintains a utility function responsible for referring a utility value U_i to each data packet following the parameter related to the packet routing. U_i acts as the expected contribution of packet i against the identified forwarding parameter. Into this utility, the packet which results in a higher increase will be replicated earlier in the RAPID protocol. RAPID will copy every message as long as the network resources, i.e., bandwidth, etc., permit [6].

5 Simulation Environment

We have analyzed the influence of the selected mobility models (both synthetic and trace-based) on the examined routing protocols on behalf of three performance metrics: delivery ratio, average latency, and overhead ratio. To do so, we have varied the values of message generation intervals (5–15 s to 45–55 s), message TTL (6 h–3 d), and buffer size (5–25 MB) successively to run simulations on the configured protocol and mobility models. When one parameter is varied, the remaining are kept fixed (message generation interval: 25–35 s, TTL: 1 d, and buffer size: 5 MB). Table 1 demonstrates the necessary parameters essential for simulations, while Table 2 represents the critical information related to selected mobility models (trace-based and synthetic). The overall simulation area is considered as 4.5×3.4 km. As the traces are campus or conference venue aligned, a single group of pedestrian is treated as mobile nodes carrying handy mobile devices. The speed of mobile nodes is assumed as 0.5–1.5 m/s. However, the number of mobile nodes is kept fixed according to the real traces. For the convenience of comparison, the same number of mobile nodes is used in the synthetic mobility model—Shortest Path Map-Based movement. Bluetooth technology is utilized as the communication interface among the mobile nodes. Due to justification and evaluating the performance comparison of simulated protocols on selected real trace-based and synthetic mobility models, we have normalized the entire simulation as 3 days. All the simulations are performed on the map of Helsinki city (the default map for ONE simulator).

Table 1 Necessary parameters for simulation [21]

Parameter	Value
Simulation area	4500 × 3400 m
Simulation time	3 d
Routing protocols	Epidemic, Spray and Wait, PRoPHET, MaxProp, and RAPID
Number of message replica (for Spray and Wait)	10
Seconds in time unit (for PRoPHET)	30
Mobility models	MIT, INFOCOM, Cambridge Imotes, and Shortest Path Map-Based
Interface type	Bluetooth
Transmit speed	2 Mbps
Transmission range	10 m
TTL	6 h, 12 h, 1d, 2 d and 3 d
Message generation intervals	5–15 s, 15–25 s, 25–35 s, 35–45 s and 45–55 s
Buffer size (MB)	5, 10, 10, 15, 20, 25
Message size (MB)	1

Table 2 Parameters related to mobility models (trace and synthetic mobility) [21]

Mobility	Type	Number of nodes	Duration
MIT reality	Trace/Campus	97	3 d
INFOCOM (Actually INFOCOM 5, also known as Haggle 3)	Trace/ Conference	41	3 d
Cambridge imotes (Haggle 4)	Trace/Campus	36	3 d
Shortest path map based	Synthetic mobility	100	3 d

6 Discussion on Simulation Outcomes

This section provides a brief discussion and analysis of simulation outcomes. We have adopted three performance metrics: delivery ratio, average latency, and overhead ratio to analyze the performance of the simulated DTN routing protocols on the considered mobility models.

Delivery Ratio: This metric indicates the ratio between the number of messages perfectly delivered to the destined targets and the total number of source-generated messages. Since delivery ratio reflects the success rate of message delivery, it is expected to be higher for a network. Figures 1, 2 and 3 demonstrate how the delivery ratio is changed with the variation of message generation intervals (5–15 s to 45–55 s), TTL (6 h–3 d), and buffer size (5–25 MB), respectively, for the selected mobility models on the simulated DTN routing protocols.

Figure 1 represents that the delivery ratio increases gradually for all mobility models on simulated protocols with the increase of message generation interval (5–15 s to 45–55 s). Here, TTL and buffer size are kept fixed at 1 d and 5 MB, respectively. Shortest Path Map-Based mobility has the highest delivery ratio than others for all protocols in this case. So, this synthetic mobility model has a higher delivery ratio than the real traces here. In addition, the limited replication strategy makes Spray and Wait as well as MaxProp the best performer for delivery ratio than other protocols, and they both exhibit a higher success rate of message delivery for the Shortest Path Map-Based Movement model. On the other hand, MIT Reality trace for Epidemic

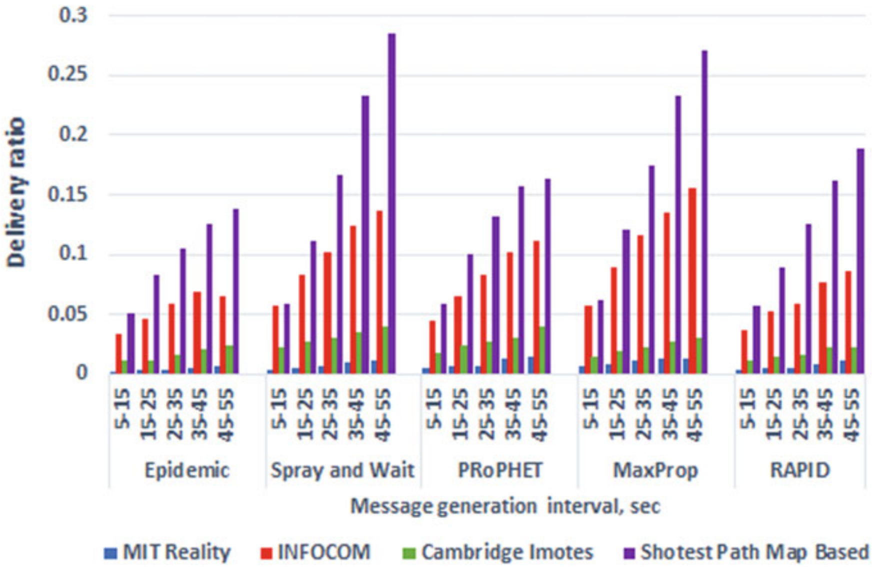


Fig. 1 Delivery ratio versus message generation interval

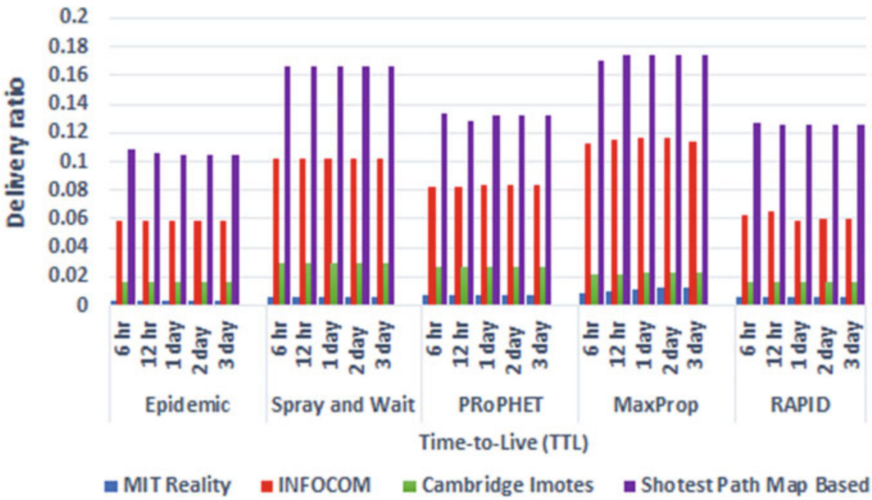


Fig. 2 Delivery ratio versus TTL

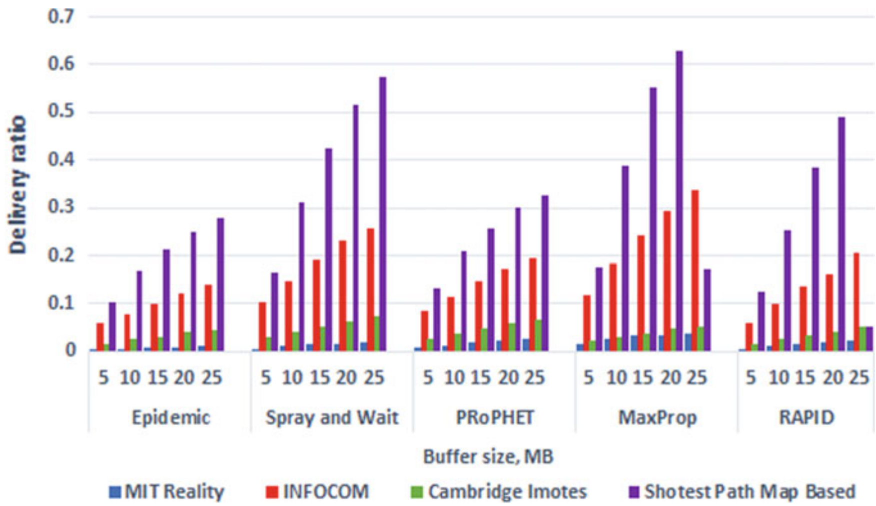


Fig. 3 Delivery ratio versus buffer size

protocol experiences the lowest value of the delivery ratio. In this case, the flooding-based forwarding strategy to route messages in Epidemic reduces the success rate of message delivery.

Figure 2 clarifies that the synthetic mobility model—Shortest Path Map-Based has the highest delivery ratio than the real traces with varying TTL (6 h–3 d). Here, message generation interval and buffer size value are maintained fixed at 25 s–35 s and 5 MB, respectively. Both Spray and Wait and MaxProp protocols have the highest delivery ratios in the Shortest Path Map-Based mobility. But between these two protocols, the MaxProp has a slightly higher delivery ratio. So, MaxProp protocol in the investigated synthetic mobility is the best performer for delivery ratio. In opposition, MIT reality for Epidemic protocol shows the worst performance for delivery ratio, similar to the prior result.

Similar to Figs. 2 and 3 provides the same result for delivery ratio with increasing buffer size (5–25 MB). Here, MaxProp protocol for Shortest Path Map-Based Movement model shows the highest delivery ratio among all protocols with real traces. Epidemic protocol for MIT Reality trace is the worst performer in this case. Here, message generation interval and TTL are kept constant at 25 s–35 s and 1 d, respectively.

Average Latency: Average delay computes the average value of the time differences among the generation of the messages (which are delivered successfully) at sources and their delivery to destination nodes. For efficient routing of messages in a network, the value of this metric requires to be small. Figures 4, 5 and 6 indicate the average latency of simulated protocols on the implemented mobility models with varying message generation intervals (5–15 s to 45–55 s), TTL (6 h–3 d), and buffer size (5–25 MB).

From Fig. 4, it is clear that the MaxProp protocol configured with MIT Reality trace has the highest latency. In contrast, INFOCOM on Spray and Wait protocol exhibits the lowest latency making it the best performer in terms of delivery delay. TTL and buffer size are set at 1 day and 5 MB, respectively, as with the previous cases.

MIT Reality for MaxProp protocol exposes the highest latency for increasing the message lifetime (6 h–3 d), while message generation interval (25–35 s) and buffer

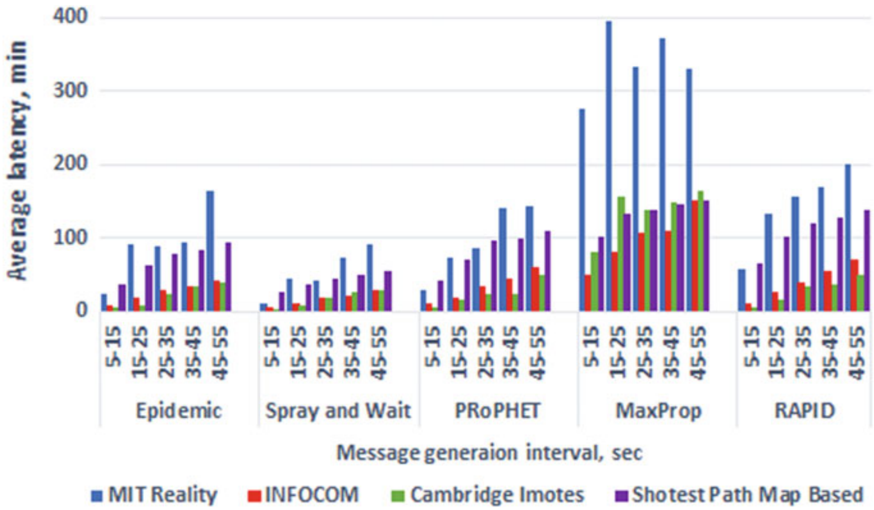


Fig. 4 Average latency versus message generation interval

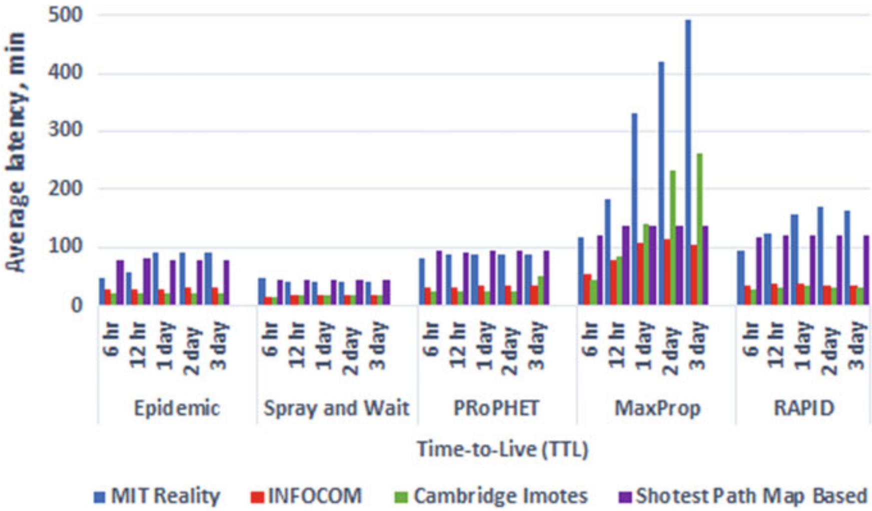


Fig. 5 Average latency versus TTL

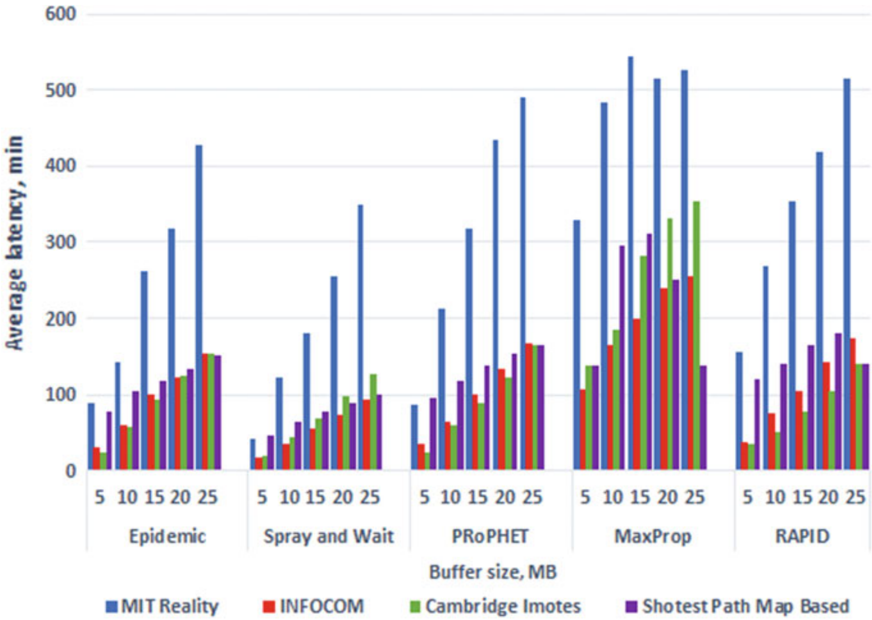


Fig. 6 Average latency versus buffer size

size (5 MB) are fixed, as in Fig. 5. So, it is the worst-performing routing protocol in this case. Furthermore, Spray and Wait protocol on Cambridge Imotes is the best performer in terms of average latency with variation in TTL.

From Fig. 6, it is observed that MIT Reality on MaxProp protocol has the highest latency value with increasing buffer size from 5 to 25 MB. So, it is the worst-performing protocol for the average latency metric. The Spray and Wait protocol configured with INFOCOM trace is the best performer since it has the lowest latency.

Overhead Ratio: This performance metric measures the transmitting efficiency of a network. It can be defined as the estimation of the number of redundant packets that are to be relayed to deliver a single data packet and are supposed to be minimum. Figures 7, 8 and 9 depict the impact of overhead ratio for the selected mobility models on the experimented protocols with varying message generation intervals (5–15 s to 45–55 s), TTL (6 h–3 d), and buffer size (5–25 MB).

It can be decided from Fig. 7 that Spray and Wait protocol configured with INFOCOM, Cambridge Imotes, and Shortest Path Map-Based Movement model has the lowest value of the overhead ratio. But among these three, INFOCOM trace has a little lesser value of the overhead ratio (for 35–45 s and 45–55 s). Except that, for other message generation interval values, these three models have the same value. So, it is evident that Spray and Wait protocol on INFOCOM is the best performer in this case. Here, TTL (1 d) and buffer size (5 MB) are maintained fixed as in the earlier. Besides, MIT Reality trace on MaxProp protocol has the highest value of

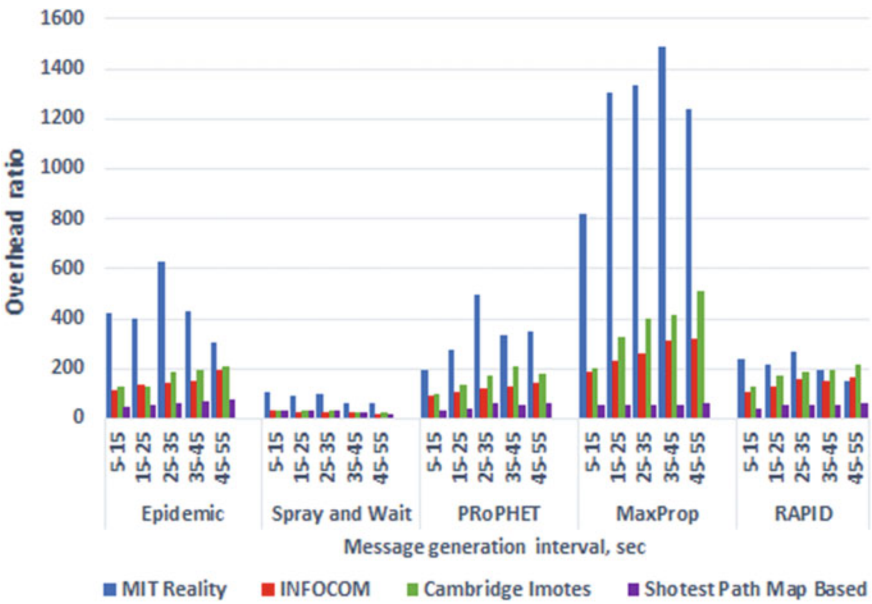


Fig. 7 Overhead ratio versus message generation interval

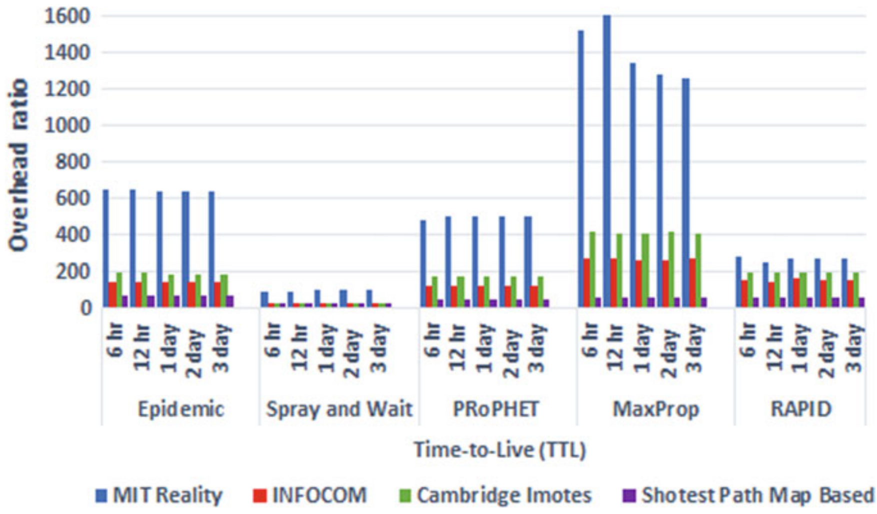


Fig. 8 Overhead ratio versus TTL

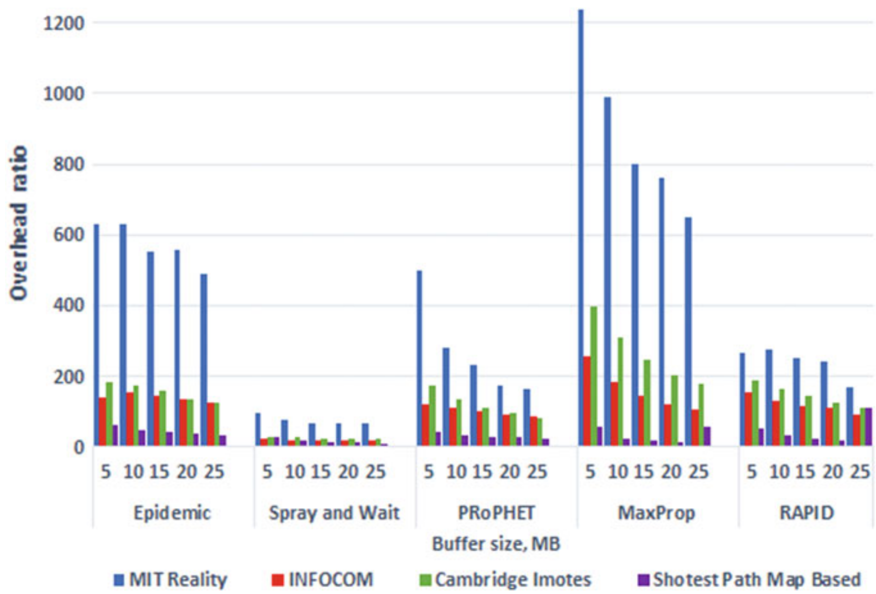


Fig. 9 Overhead ratio versus buffer size

overhead ratio with the increase of message generation interval. So, it is the worst performer in this case.

MIT Reality for MaxProp protocol exploits the maximum overhead ratio in Fig. 8 among the others with the variation of TTL from 6 h to 3 days, while message generation interval (25–35 s) and buffer size (5 MB) are kept constant. On the contrary, it seems that INFOCOM, Cambridge Imotes, and Shortest Path Map-Based Movement Model for Spray and Wait protocol have the lowest overhead ratio values altogether. But among the three mobility models, INFOCOM trace configured with Spray and Wait protocol has the least minimum values of the overhead ratio. So, INFOCOM trace is the best-performing trace on Spray and Wait protocol.

It can be seen from Fig. 9 that the worst performer for the overhead ratio is MIT reality in MaxProp protocol against the change of buffer size. Message generation interval and TTL are preserved at 25–35 s and 1 day, respectively. Contrarily, Shortest Path Map-Based Movement model on Spray and Wait performs the best compared to all since the value of the overhead ratio is the lowest for it in such a case.

7 Conclusions and Future Work

In this research, we have analyzed the impacts of the examined mobility models (three trace-based mobility models and one synthetic mobility model) on the performance of the five DTN routing protocols mentioned above, considering the variations of message generation intervals, TTL, and buffer size in turn. After evaluating all of the simulation outcomes, we conclude that Shortest Path Map-Based Movement model on MaxProp protocol has the highest delivery ratio with the variation of both TTL and buffer size. Furthermore, this synthetic mobility model has the highest delivery ratio for the increasing message generation interval for MaxProp and Spray and Wait protocols. On the other hand, MIT Reality on the Epidemic protocol is the worst performing for delivery ratio with the change of message generation interval, TTL and buffer size, respectively. Besides, MIT Reality on MaxProp protocol shows the worst performance for average latency and overhead ratio with varying message generation interval, TTL and buffer size. In addition, INFOCOM on Spray and Wait protocol for average latency and overhead ratio with varying message generation intervals, Cambridge Imotes on Spray and Wait with varying TTL for average latency, and INFOCOM on Spray and Wait for overhead ratio with varying TTL exhibit the best performance. Furthermore, Spray and Wait on INFOCOM is the best performer for average latency with varying buffer sizes. At the same time, Shortest Path Map-Based Movement on Spray and Wait has the best performance for overhead ratio. The future research endeavor will investigate the impacts of trace-based mobility models on the performance and energy consumption of social-aware DTN routing protocols from different perspectives.

Acknowledgements This work has been supported by National Science and Technology (NST) MPhil fellowship, granted by Ministry of Science and Technology, Government of the People's Republic of Bangladesh.

References

1. Ali, S., Qadir, J., Baig, A.: Routing protocols in delay tolerant networks—a survey. In: Proceedings—2010 6th International Conference Emerging Technology ICET 2010. pp. 70–75 (2010). https://doi.org/10.1109/ICET.2010.5638377.
2. Roy, A., Acharya, T., DasBit, S.: Quality of service in delay tolerant networks: a survey. Comput. Netw. **130**, 121–133 (2018). https://doi.org/10.1016/j.comnet.2017.11.010
3. Vahdat, A., Becker, D.: Epidemic routing for partially-connected ad hoc networks. Department of Computer Science, Duke University, Tech. Rep. CS–2000–06 (2000)
4. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In: Proceedings ACM SIGCOMM 2005 Workshop Delay-Tolerant Networking, WDTN 2005. pp. 252–259. (2005). https://doi.org/10.1145/1080139.1080143.
5. Lindgren, A., Doria, A., Schelén, O.: Probabilistic routing in intermittently connected networks. ACM Mobile Comput. Commun. Rev. **7**(3), 19–20 (2003). https://doi.org/10.1145/961268.961272
6. Balasubramanian, A., Levine, B.N., Venkataramani, A.: DTN routing as a resource allocation problem. In: SIGCOMM, Kyoto, Japan, pp. 373−384. ACM (2007). https://doi.org/10.1145/1282380.1282422.
7. Burgess, J., Gallagher, B., Jensen, D., Levine, B.N.: MaxProp: routing for vehicle-based disruption-tolerant networks. In: IEEE INFOCOM, Barcelona, Spain, pp. 1−11. IEEE (2006). https://doi.org/10.1109/INFOCOM.2006.228.
8. Keränen, A., Ott. J.: Increasing reality for dtn protocol simulations. Helsinki University of Technology, Tech. Rep. (2007)
9. Keränen, A., Kärkkäinen, T., Ott, J.: Simulating mobility and dtns with the one. J. Commun. **5**(2), 92–105 (2010). https://doi.org/10.4304/jcm.5.2.92-105
10. A community resource for archiving wireless data at dartmouth, crawdad, (2013) http://crawdad.cs.dartmouth.edu/. Accessed 30 Sep 2021, 1:05:20 AM, Bangladesh Standard Time
11. Mota, V.F.S., Cunha, F.D., Macedo, D.F., Nogueira, J.M.S., Loureiro, A.A.F.: Protocols, mobility models and tools in opportunistic networks: a survey. Comput. Commun. **48**, 5–19 (2014). https://doi.org/10.1016/j.comcom.2014.03.019
12. Keränen, A., Ott, J., Kärkkäinen, T.: The one simulator for dtn protocol evaluation. In: SIMUTools 2009–2nd International ICST Conference Simulation Tools Technolgy (2009). https://doi.org/10.4108/ICST.SIMUTOOLS2009.5674.
13. Socievole, A., De Rango, F., Coscarella, C.: Routing approaches and performance evaluation in delay tolerant networks. In: Wireless Telecommunications Symposium (2011). https://doi.org/10.1109/WTS.2011.5960870
14. Cabacas, R., Ra, I.H.: Evaluating mobility models in delay tolerant network. In: 2013 International Conference IT Converging Security ICITCS 2013. pp. 5–8 (2013). https://doi.org/10.1109/ICITCS.2013.6717832.
15. Bista, B.B., Rawat, D.B.: Energy consumption and performance of delay tolerant network routing protocols under different mobility models. In: ISMS, pp. 325−330. IEEE (2016). https://doi.org/10.1109/ISMS.2016.21.
16. Spaho, E., Bylykbashi, K., Barolli, L., Kolici, V., Lala, A.: Evaluation of different dtn routing protocols in an opportunistic network considering many-to-one communication scenario. In: NBiS 2016—19th International Conference Network-Based Information Systems, pp. 64–69. (2016). https://doi.org/10.1109/NBiS.2016.84.

17. Khan, M.K.M., Rahim, M.S.: Performance analysis of social-aware routing protocols in delay tolerant networks. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). pp. 1–4. IEEE (2018). Rajshahi, Bangladesh (2018). https://doi.org/10.1109/IC4ME2.2018.8465592.
18. Spaho, E., Dhoska, K., Bylykbashi, K., Barolli, L., Kolici, V., Takizawa, M.: Performance evaluation of routing protocols in dtns considering different mobility models. In: Workshops of the International Conference on Advanced Information Networking and Applications, pp. 205−214. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15035-8_19.
19. Khan, M.K.M., Roy, S.C., Rahim, M.S., Islam, A.Z.M.T.: On the energy efficiency and performance of delay-tolerant routing protocols. In: Bhuiyan, T., Ali, M. (eds.) Cyber Security and Computer Science. ICONCS 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunication Engineering, vol. 325, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52856-0_44.
20. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Pers. Ubiquitous Comput. **10**, 255–268 (2006). https://doi.org/10.1007/s00779-005-0046-3
21. Khan, M.K.M., Rahim, M.S., Islam, A.Z.M.T.: Impact of trace-based mobility models on the energy consumption of delay-tolerant routing protocols. In: 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). pp. 1–6. IEEE (2021). Dhaka. https://doi.org/10.1109/ICEEICT53905.2021.9667942.

A Study on DC Characteristics of Si-, Ge- and SiC-Based MOSFETs

Kazi Mahamud Al Masum, Tanvir Yousuf Shohag, and Md. Shahid Ullah

Abstract This paper instigates the findings on switching and DC characteristics such as transfer characteristics and output characteristics as well as carrier concentration, electric potentials of the metal–oxide–semiconductor field-effect transistors (MOSFET) after changing its materials by silicon (Si), germanium (Ge) and silicon carbide (SiC) using COMSOL Multiphysics 5.5 software. Though there have been various researches on the MOSFET with many materials for years, the comparative study on MOSFET's DC characteristics is presented here showing what changes happened with the materials of it. Among the three nMOS, the Ge-MOSFET notably shows the minimum threshold voltage along with maximum terminal current rating than the other Si- and SiC-based MOSFETs for exactly the same study setup, which means the Ge-based MOSFET as a voltage-controlled device needs less voltage to switch but give a higher range of saturation current as a current source, which all are discussed here.

Keywords MOSFET · Power device · DC characteristics · Switching · Semiconductor device · Si · Ge · SiC · Wide bandgap material

1 Introduction

The trend of power electronics devices is run based on the operation in high voltage, the high temperature as well as high density, and after the age of Si, as a wide-bandgap semiconductor material, SiC started playing a crucial but popular role in the evolution process of this industry for the last two decades and Ge also becoming

K. M. Al Masum (✉) · T. Y. Shohag · Md. S. Ullah
Department of Electrical and Electronic Engineering, International Islamic University Chittagong, Chattogram 4000, Bangladesh
e-mail: mahamud.masum123@gmail.com

T. Y. Shohag
e-mail: tanvir.iiuc161027@gmail.com

Md. S. Ullah
e-mail: Shahideee04@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_24

popular in some aspects [1–4]. Another critical aspect of this topology is the switching characteristics of power switches like power diodes, MOSFET, IGBT and BJT [1–9].

Vast research has been going on power MOSFET and its overall performance with different materials and structures of MOSFET used [1, 2, 10–13]. SiC devices show about 80% lower loss in switching than the Si devices in some aspects so that hybrid application of Si-IGBT and SiC-MOSFET was also proposed [14]. On the aspect of on-resistance and high-temperature application, the SiC-MOSFET showed better performance than the Si-MOSFET [5]. For overcoming MOSFET's challenges, research on junction field-effect transistors (JFET) was done by using different materials like Si, SiC and others [15]. Researchers presented how the MOSFET's oxide layer can be degraded by the total ionization dose causing the change in threshold voltage and current rating [16]. DC characteristics of a MOSFET under optical illumination were also researched [16]. Most of these types of research are based on Si, SiC, GaAs, GaN, etc., but the other materials like Ge [15].

However, the DC characteristics of MOSFET for specific parameter settings are discussed in this paper after varying the semiconductor materials by Si, Ge and SiC. The responses of the three MOSFETs based on Si, Ge and SiC sequentially in terms of the electron concentration, hole concentration, electric potentials, transfer characteristics and output characteristics by using the COMSOL Multiphysics simulation software. Throughout the analysis, it's observed that for certain methodology the Ge-MOSFET shows better responses than the other two despite the triumph of Si and SiC [2, 5, 6].

2 MOSFET's Basic Concept

2.1 NMOS Construction

A semiconductor device, MOSFET is one type of field-effect transistors, which has oxide insulation in the gate metal. Among many types of MOSFET, the enhancement type MOSFET with n-channel is more popular because of its advantages over others and its construction is shown in Fig. 1 [13, 16].

An nMOS is a voltage-controlled device with highly N-Type doped source and drain region as well as p-type doped body region, and it has four types of terminals, which are known as source(S), drain(D), gate(G) and body(B) [13].

2.2 N-Type MOSFET's Switching and Its DC Characteristics

Since the source and body both are being grounded, then the gate-to-source voltage (V_{gs}) is responsible for creating the channel between the source to drain as well as controlling the threshold voltage (V_t), which is the cause the channel to start creating,

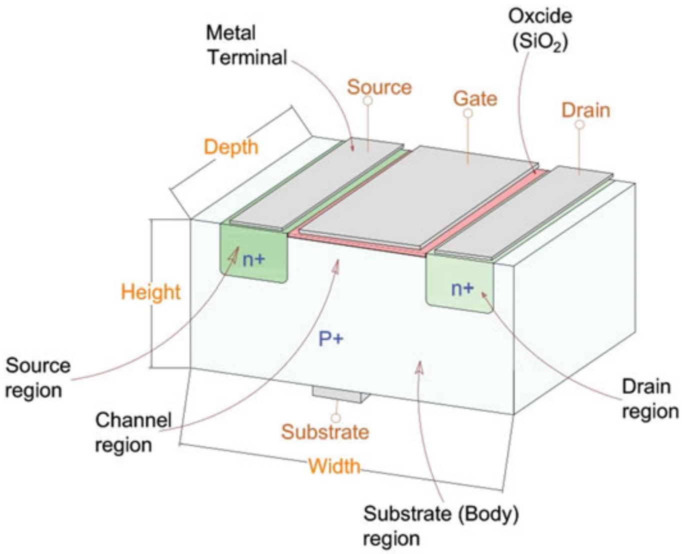


Fig. 1 An N-Type enhancement MOSFET from the perspective view [16]

and finally, the drain-to-source voltage (V_{ds}) is controlled the terminal current (I_d) [4]. The output characteristics curve with the triode and saturation region of a basic nMOS is shown in Fig. 2 [16, 17].

3 Experimental Setups

3.1 *The Geometry and Mesh of the Experimental N-Type MOSFET*

The geometry of the MOSFET of width 3 µm, depth 0.1 µm and height 0.7 µm is shown in Fig. 3a, and the mesh of the model is presented in Fig. 3b, respectively. This same geometry and mesh of Fig. 3 of the MOSFET are used for all the three Si-, Ge- and SiC-based MOSFETs in this paper.

As a baSiC-MOSFET structure shown in Fig. 1, the N^+ Source, as well as the N^+ Drain regions both are 0.03 µm long with 0.5 µm width along with N-Type doping of $1 \times 10^{20} \frac{1}{\text{cm}^3}$. All other portion is considered as the body region which is p-type doped with $1 \times 10^{17} \frac{1}{\text{cm}^3}$. The oxide layer of the gate is 30 nm with the relative permittivity of 4.5. The simulation is set up with the default temperature of 293.15 K.

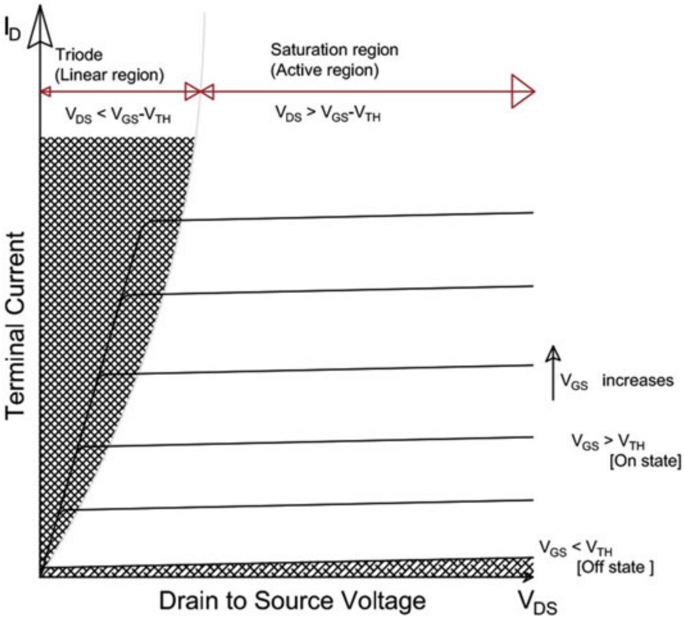


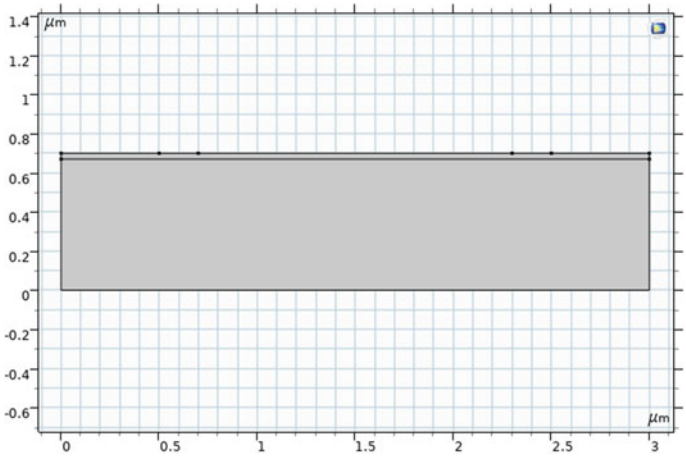
Fig. 2 Output characteristics of an enhancement N-Type MOSFET [17]

3.2 The Materials and Its Properties Selection

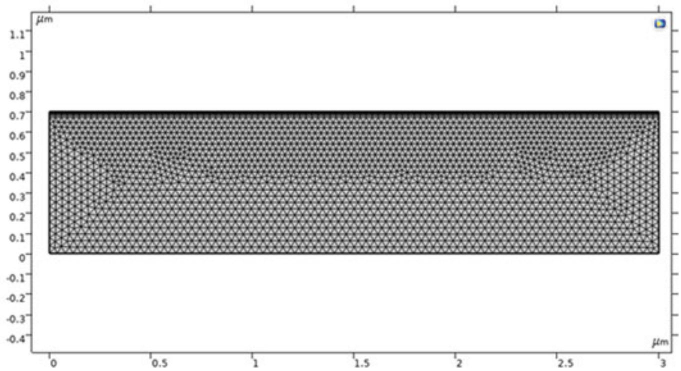
In Table 1, the material properties applied for Si, Ge, SiC are documented. The most basic properties are included here and others are generated by COMSOL Multiphysics simulation software.

3.3 The Study Setups

The study is set up with gate voltage (V_g) for a range 1–5 V with an interval of 0.5 V as well as the drain voltage (V_d) for a range 0–5 V with an interval of 0.5 V to observe the transfer characteristics, output characteristics and others, which are represented for all three MOSFETs in the following.



(a) Geometry of the MOSFETs



(b) Mesh of the MOSFETs

Fig. 3 Structural design used for all three materials MOSFETs

4 Results and Discussions

4.1 Transfer Characteristics

Figure 4 shows the transfer characteristics for the Si, Ge and SiC-MOSFETs, respectively. In this result, it is shown that the Ge-MOSFET gets an advantage as a switching device as it needs the minimum threshold voltage (V_t) about 1 V, whereas the Si-MOSFET needs 1.2 V and the SiC-MOSFET needs 2.8 V approximately.

Since the SiC-MOSFET needs a higher threshold voltage, for the drain voltage below 2.8 V, this cannot be operated as well as the output characteristics curve cannot be found for this range.

Table 1 Applied properties of the three materials [3, 5, 18, 19]

Property	Applied values for		
	Si	Ge	SiC
Relative Permittivity	11.7	16.2	9.7
Thermal conductivity [W/m K]	131	60	450
Density [kg/m^3]	2329	5323	3200
Band gap [eV]	1.12	0.664	3.26
Electron affinity [eV]	4.05	4	3.24
Effective density of state, the conduction band [1/m^3] [1/m^3]	2.8×10^{19}	1.04×10^{19}	1.7×10^{25}
Effective density of state, the valence band [1/m^3]	1.04×10^{19}	6.0×10^{18}	2.5×10^{25}
Mobility of the electron [m^2/$V\ s$]	0.145	0.390	0.090
Mobility of the hole [m^2/$V\ s$]	0.050	0.190	0.012
Electron lifetime, SRH [µs]	10	1×10^{-3}	2.5×10^{-6}
Hole lifetime, SRH [µs]	10	1×10^{-4}	5.0×10^{-7}

4.2 Electron Concentrations

Figure 5 shows the electron concentrations for the Si-, Ge- and SiC-based MOSFETs, respectively. Among all surface results for the given values of Vd with a fixed Vg, only the last one result of the range, which is for $Vd = 5$ V and $Vg = 4$ V is shown in Fig. 5 for all the three MOSFETs, respectively.

However, for the same setup and same drain and gate voltage, the SiC-MOSFET has a larger depletion region than the other two which is found in both Figs. 5c and 6c, which may affect output characteristics. The material properties of SiC may be responsible for this characteristic.

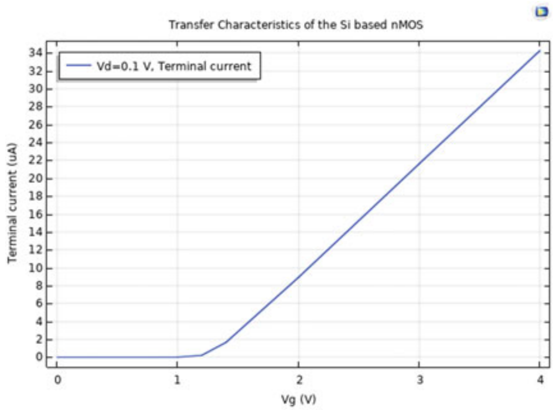
4.3 Hole Concentrations

Figure 6 shows the hole concentration for the Si-, Ge- and SiC-based MOSFETs, respectively. In Fig. 6, we observe how the hole concentrated more around the source, channel and drain side are less down with the increase of V_g and V_d.

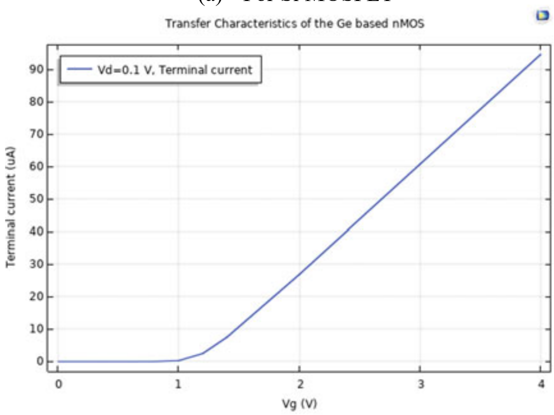
The depletion region is also increased with the increase of V_g and V_d, which is clearly noticed by Fig. 6 for all the three Si, Ge and SiC-MOSFET, respectively. Similar to Fig. 5c, the SiC-MOSFET shows a larger depletion region than Si-MOSFET and Ge-MOSFET.

The depletion region is also increased with the increase of V_g and V_d, which is clearly noticed by Fig. 6 for all the three Si, Ge and SiC-MOSFET, respectively. Similar to Fig. 5c, the SiC-MOSFET shows a larger depletion region than the Si-MOSFET and Ge-MOSFET.

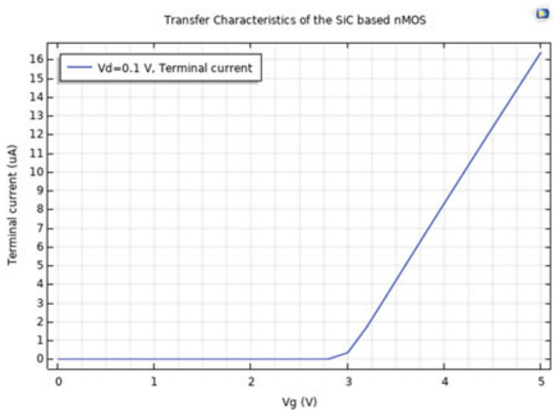
Fig. 4 Transfer characteristics of Si, Ge and SiC-MOSFETs



(a) For Si MOSFET

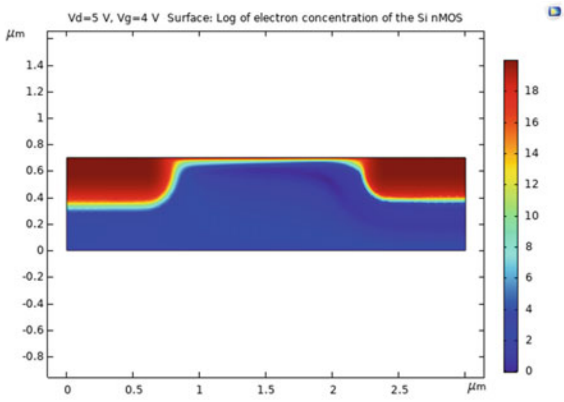


(b) For Ge MOSFET

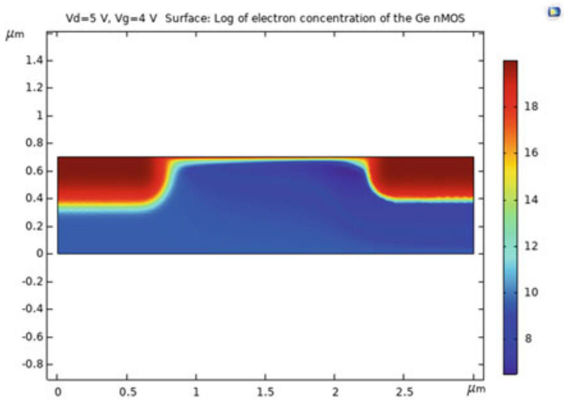


(b) For SiC MOSFET

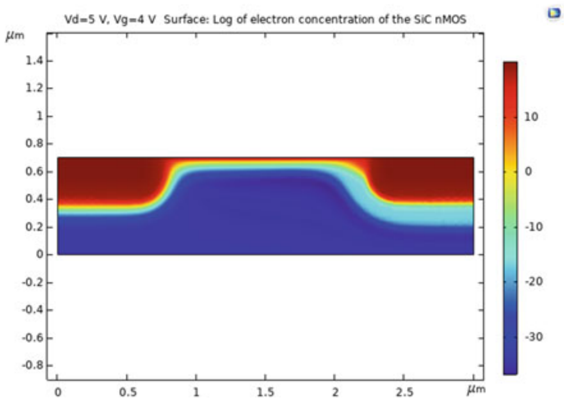
Fig. 5 Transfer characteristics of Si, Ge and SiC-MOSFETs



(b) For Si MOSFET

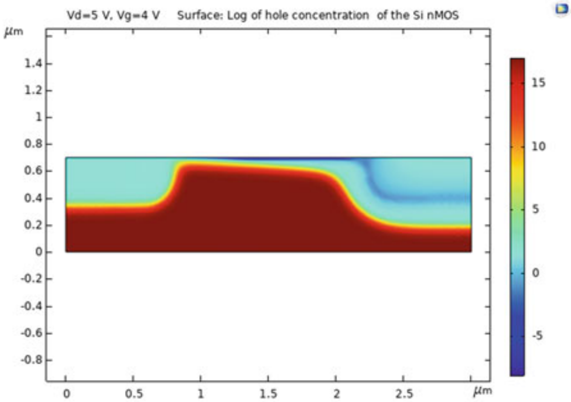


(b) For Ge MOSFET

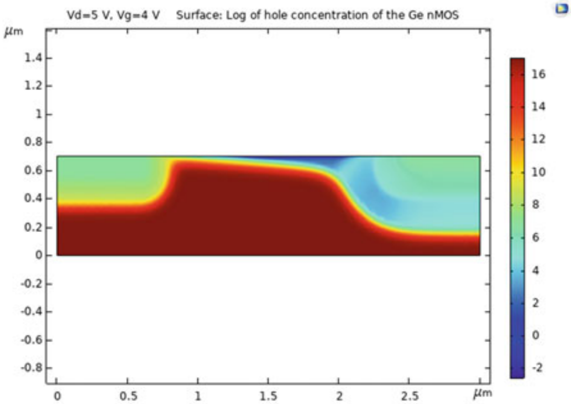


(b) For SiC MOSFET

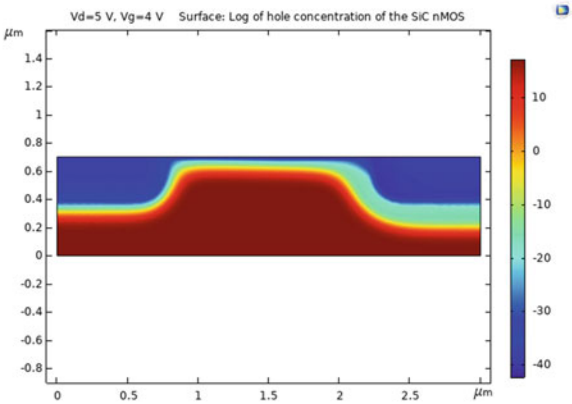
Fig. 6 Hole concentrations of Si, Ge and SiC-MOSFETs



(a) For Si MOSFET



(b) For Ge MOSFET



(c) For SiC MOSFET

4.4 Electric Potentials

Figure 7 shows the electric potentials for the Si-, Ge- and SiC-based MOSFETs, respectively. For all the three MOSFETs, the drain has the highest electric potentials as the MOSFETs are reached saturation, which is expected due to the given value of $V_d \geq (V_g - V_t)$, discussed in Fig. 2.

4.5 Output Characteristics

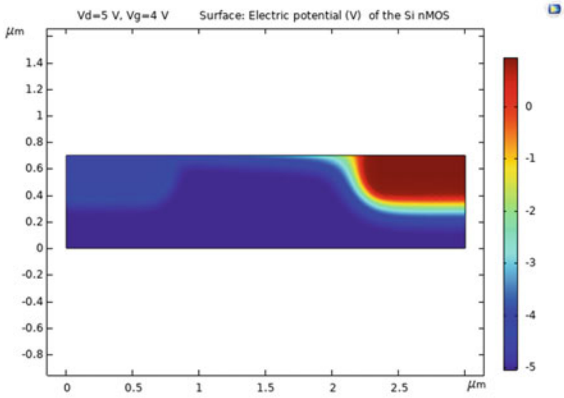
Figure 8 shows the output characteristics for the Si, Ge and SiC-MOSFETs, respectively. For Si MOSFET, since $V_t = 1.2$ V, is shown in Fig. 4a, we see the output curves for different values of V_g, the terminal currents of the saturation region are different. For example, for $V_g = 5$ V, we get the terminal current of saturation is about 700 μ A, which is shown in Fig. 8a.

After that, for Ge-MOSFET, since $V_t = 1.0$ V, is shown in Fig. 4b, we see the output curves for different values of Vg, the terminal currents of the saturation region are different. For example, for $V_g = 5$ V, we get the terminal current of saturation is higher than the other Si and SiC-MOSFETs, about 1800 μA, which is shown in Fig. 8b.

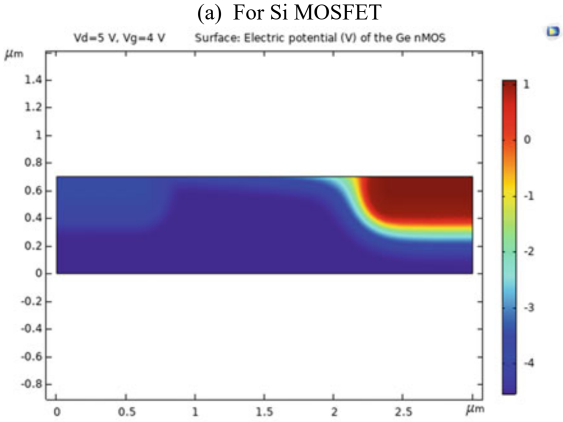
Finally, due to the SiC-MOSFET's $V_t = 2.8$ V, found in Fig. 4c, the output curves couldn't be found for the range of $1-2.5$ V of V_g since the channel can't be created. In Fig. 8c, after the $V_g = 2.8$ V, the output curves are presented, though the terminal current of the saturation is much lower than the Si- and Ge-based MOSFETs for the same parameters and values. In this study, most of the material properties of SiC are close to the 4H-SiC so that for other types of SiC like 6H-SiC and 3H-SiC the result may be different.

Another important factor in this study is the simulation and is run with the COMSOL Multiphysics 5.5 software which is itself run with pre-assumed and predetermined values and parameters, which could affect the whole simulation process. However, the geometry of MOSFET used for all three times plays an important role in the characteristics. By changing the oxide layers, doping concentrations, other parameters of the structure or properties of materials, the characteristics would be dramatically changed. But for any study or operational setup like the above scenarios, the Ge-based MOSFET can give better performance, which may show us an alternative, maybe the better scope over today's Si material and devices. Moreover, there are many other research scopes on this particular topic to get faster and efficient semiconductor devices.

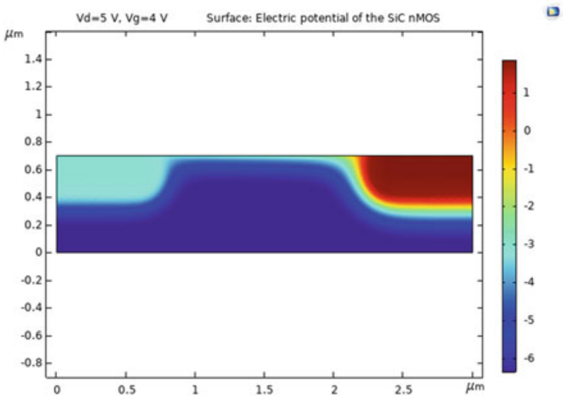
Fig. 7 Electric potentials of Si, Ge and SiC-MOSFETs



(a) For Si MOSFET

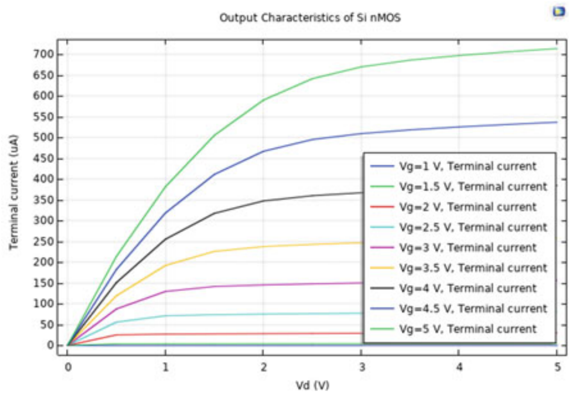


(b) For Ge MOSFET

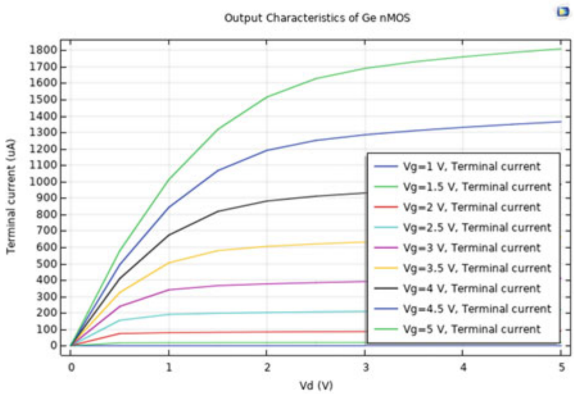


(c) For SiC MOSFET

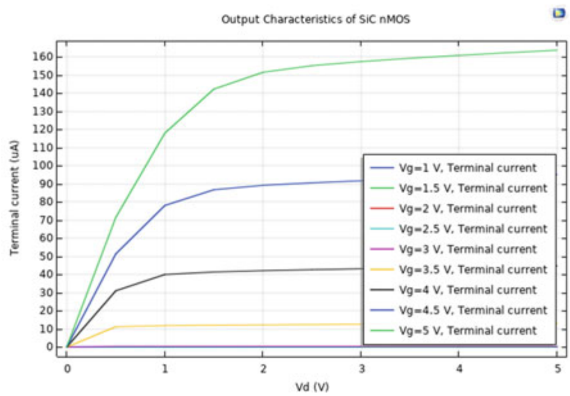
Fig. 8 Output characteristics of Si, Ge and SiC-MOSFETs



(a) For Si MOSFET



(b) For Ge MOSFET



(c) For SiC MOSFET

5 Conclusion

The results for the three different materials MOSFETs, found by the simulations, are surprising as the Ge-MOSFET has shown better switching response than the other Si and SiC-MOSFETs on the basis of lower threshold voltage and higher current rating. The Ge-MOSFET needs half the threshold voltage than the SiC-MOSFET for this setup and more than two times more terminal current than Si-MOSFET and more than ten times more terminal current than SiC-MOSFET. Though this study isn't challenging for migrating to other semiconductor devices from Si material, this indicates the possibility and positivity of future research not only on the SiC devices but also on other semiconductors like Ge devices.

References

1. She, X., Huang, A.Q., Lucia, O., Ozpineci, B.: Review of silicon carbide power devices and their applications. IEEE Trans. Ind. Electron. **64**, 8193–8205 (2017). https://doi.org/10.1109/TIE.2017.2652401
2. Millan, J., Godignon, P., Perpina, X., et al.: A survey of wide bandgap power semiconductor devices. IEEE Trans. Power Electron. **29**, 2155–2163 (2014). https://doi.org/10.1109/TPEL.2013.2268900
3. Houssa, M., Satta, A., Simoen, E., et al.: In: Electrical Performance of Ge Devices. Elsevier Ltd (2007)
4. Neamen, D.A.: In: Semiconductor Physics and Devices: Basics Applications (2012)
5. Wang, J.: A comparison between Si and SiC-MOSFETs. IOP Conf Ser Mater Sci Eng **729** (2020).https://doi.org/10.1088/1757-899X/729/1/012005
6. Van Wyk, J.D., Lee, F.C.: On a future for power electronics. IEEE J. Emerg. Sel. Top Power Electron. **1**, 59–72 (2013). https://doi.org/10.1109/JESTPE.2013.2271499
7. Kassakian, J.G., Jahns, T.M.: Evolving and emerging applications of power electronics in systems. IEEE J. Emerg. Sel. Top Power Electron. **1**, 47–58 (2013). https://doi.org/10.1109/JESTPE.2013.2271111
8. Wang, F., Wang, G., Huang, A., et al.: Design and operation of A 3.6kV high performance solid state transformer based on 13kV SiC-MOSFET and JBS diode. In: 2014 IEEE Energy Conversation Congress Exposition ECCE 2014, pp. 4553–4560. (2014) https://doi.org/10.1109/ECCE.2014.6954024
9. Flack, T.J., Pushpakaran, B.N., Bayne, S.B.: GaN technology for power electronic applications: a review. J. Electron. Mater. **45**, 2673–2682 (2016). https://doi.org/10.1007/s11664-016-4435-3
10. Lee, C.H., Nishimura, T., Tabata, T., et al.: Ge MOSFETs performance: impact of Ge interface passivation. Tech Dig—Int Electron Devices Meet IEDM **2**, 416–419 (2010). https://doi.org/10.1109/IEDM.2010.5703384
11. Ghibaudo, G., Chroboczek, J.: On the origin of the LF noise in Si/Ge MOSFETs. Solid State Electron. **46**, 393–398 (2002). https://doi.org/10.1016/S0038-1101(01)00112-5
12. Kibushi, R., Hatakeyama, T., Yuki, K., et al.: SiC Power MOSFET using Electro-Thermal Analysis. pp. 188–192. (2020)
13. Zabeli, M., Caka, N., Limani, M., Kabashi, Q.: The impact of MOSFET's physical parameters on its threshold voltage. In: MINO'07 Proceedings 6th Conferene Microelectronics Nanoelectronics Optoelectronics, pp. 54–58. (2007)
14. Nawaz, M., Ilves, K.: Replacing Si to SiC: opportunities and challenges. In: European Solid-State Device Research Conference 2016-October, pp. 472–475. (2016). https://doi.org/10.1109/ESSDERC.2016.7599688

15. Gu, X., Shui, Q., Myles, C.W., Gundersen, M.A.: Comparison of Si, GaAs, SiC and GaN FET-type switches for pulsed power applications. In; Digest Technical Paper International Pulsed Power Conference, pp. 362–365. (2003).https://doi.org/10.1109/ppc.2003.1277729
16. Kannan, R., Khalid, H., Indragandhi, V., Albert Alexander, S.: Threshold voltage and drain current investigation of power MOSFET ZVN3320FTA by 2D simulations. Int. J. Simul. Syst. Sci. Technol. 19, 39.1–39.6 (2018). https://doi.org/10.5013/IJSSST.a.19.06.39
17. Jain, P., Mishra, K.B.: An investigation of DC characteristics in multifinger optically illuminated MOSFET. Int. J. Comput. Appl. **61**, 12–17 (2013). https://doi.org/10.5120/9899-4479
18. Chapter 5—Field-Effect Transistors (FETs)—ppt video online download. https://slideplayer.com/slide/5122876/. Accessed 27 Nov 2021
19. Analysis of MOSFET as a Switch with Circuit Diagram, Example. https://www.electronicshub.org/mosfet-as-a-switch/. Accessed 27 Nov 2021
20. Kimoto, T.: SiC material properties. Elsevier Ltd (2018)
21. Kazimierczuk, M.K.: Physical constants. Phys. Rev. D **54**, 65–67 (1996). https://doi.org/10.2307/j.ctv1fkgcmf.7

A Compact UWB Array Antenna for Microwave Imaging/WiMAX/Wi-Fi/Sub-6 GHz Applications

Liton Chandra Paul, Golam Arman Shaki, Tithi Rani, and Wang-Sang Lee

Abstract This article represents the design and performance analysis of a compact ($36.25 \times 34.75 \times 0.79$ mm^3) ultra-wideband (UWB) microstrip patch array antenna having parasitic elements whose operating frequency ranges from 2.77 to 12.85 GHz. The UWB antenna is mounted on Roger's RT 5880 (lossy) substrate (2.2, 0.0009). The entire design process and simulation is performed by CST. The results of the designed UWB antenna show that the reflection coefficient is −32.8 dB at 3.45 GHz, −35.33 dB at 4.68 GHz, −22.524 dB at 7.78 GHz, and 18.23 dB at 12.09 GHz, gain is 3.92 dB at 4.68 GHz, directivity is 4.55 dBi at 4.68 GHz, and VSWR is 1.0348 at 4.68 GHz. Moreover, the maximum radiation efficiency is 98%. Due to compact size (995.153 mm^3) and ultra wide operating band of 10.08 GHz, the designed antenna is a suitable model for many wireless applications including microwave imaging (3.1–10.6 GHz), WiMAX (3.4–3.6 GHz), WLAN, WiFi-5/6 (5.180−5.925 GHz, 5.940–7.075 GHz), and Sub-6 GHz (3.3–4.9 GHz) applications. The designed antenna has a stable radiation pattern with a good gain and efficiency.

Keywords Microstrip patch array · UWB · Microwave imaging · WLAN · WiFi-5/6 · WiMAX · Sub-6 GHz · CST

L. C. Paul (✉) · G. A. Shaki
Department of EECE, Pabna University of Science and Technology, Pabna 6600, Bangladesh
e-mail: litonpaulete@gmail.com

G. A. Shaki
e-mail: shakiarman09@gmail.com

T. Rani
Department of ETE, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh
e-mail: tithi.saha.134040@gmail.com

W.-S. Lee
Department of Electronic Engineering, Gyeongsang National University, Jinju 52828, South Korea
e-mail: wsang@gnu.ac.kr

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_25

1 Introduction

An antenna is a microwave transducer because it converts electrical signal to electromagnetic signal or radio waves which deals with microwave frequencies (300−300 GHz) and vice versa. In this modern age, the revolution in communication network technology requires the enlargement of low profile antennas that are efficient in holding high performance over a wide range of frequency [1]. It is noted that in the last two decades, ultra-wideband communication has achieved extraordinary consideration due to its potential facilities like low power utilization, high energy radiation, the rapid data transmission rate, low power density, and larger frequency spectrum which delivers an excellent output to the consumers. The ultra-wideband (UWB) can be spelled out as the technology of transferring data over a wide bandwidth (BW) along with the transmission of −10 dB BW which is more than one-fourth or 25% of a center frequency [2, 3]. The military radar and satellite communication systems require ultra-wideband antennas [4]. Microstrip patch antennas (MPAs) serve the purpose for such applications as they are low profile, easy to fabricate, occupy less volume, and mechanically robust when applied on rigid surfaces [5].

An UWB system anatomy is defined by the Federal Communications Commission (FCC) where any radio system operates within 3.1−10.6 GHz [6, 7]. Moreover, it is observed that ultra-wideband (UWB) antennas have a stable radiation pattern, high impedance matching, and high performance. The UWB communication system has endured appreciable recognition from both industrial and academic sectors [8]. Generally, antennas are designed to radiate the power in the desired pattern. Though numerous UWB antennas are presented and designed by different researchers in different shapes like rectangular shaped, elliptical shaped, pentagon shaped, hexagonal shaped, U-shaped, C-shaped, lamp post shaped, circular shaped, square shaped, etc., the designers are still analyzing UWB for further development. However, the maximum gain is not high enough for many published UWB patch antennas, especially for compact patch antennas. In [9], the square-shaped antenna is mounted on FR-4 substrate having an operating frequency band of 2.33−12.4 GHz. Though efficiency is greater than 90%, the size of the antenna is comparatively large (40 × 43 mm^2). Authors have presented a six circular cut hexagonal-shaped antenna (36 × 36 × 1.6 mm^3) designed on FR-4 [10]. The operating frequency band is very large (3−27.57 GHz) with a percentage bandwidth of 160.75%. A circularly polarized antenna is discussed in [11], where the dimension of the designed antenna is 60 × 55 × 1.59 mm^3 building on FR-4 substrate ($\varepsilon_r = 4.4$). This UWB antenna resonates at 4.3, 5, 6.1, 7.4, 8.9, and 9.2 GHz having gains of 1.08, 3.23, 3.36, 2.77, 3.07, and 4.87 dB. A heart-shaped antenna is presented in [12] for UWB applications whose dimension is 0.2λ × 0.17λ covering the frequency band of 2.90−10.70 GHz. The gain and radiation efficiency of the designed antenna are 5.3 dB and 86.6%, respectively. The authors have claimed this antenna is applicable for radar and microwave imaging. In [13], a U-shaped dielectric resonator antenna (DRA) is presented which operates within the band of 2.4−11.4 GHz.

A compact microstrip patch array antenna (CMPAA) is proposed for UWB applications operating within a band from 2.77 to 12.85 GHz with a percentage bandwidth (PBW) of 129% which satisfies the UWB spectrum determined by FCC (3.1−10.6 GHz). The UWB CPMAA having a large bandwidth of 10.08 GHz is suitable for microwave imaging, WiMAX (3.4−3.6 GHz), lower 5G (3.33−4.2 GHz), 5 GHz WiFi (5.180−5.925 GHz), 6 GHz WiFi (5.940–7.075 GHz), Sub-6 GHz as well as X-band (8–12 GHz). The CPMAA is also capable of operating for both uplink (5.925−6.425 GHz) and downlink (3.7−4.2 GHz) in satellite communication. Simulation, analysis, and optimization of this proposed antenna are performed by using the simulation tool named Computer Simulation Technology (CST Studio Suite). Section 2 represents the design and structure of the antenna. In Sect. 3, different simulated results are represented such as reflection coefficient graph, 2D, 3D, and Cartesian plot of gain and directivity, E and H-fields, current distribution, Z-parameter, and VSWR. And at last, conclusion is presented in Sect. 4.

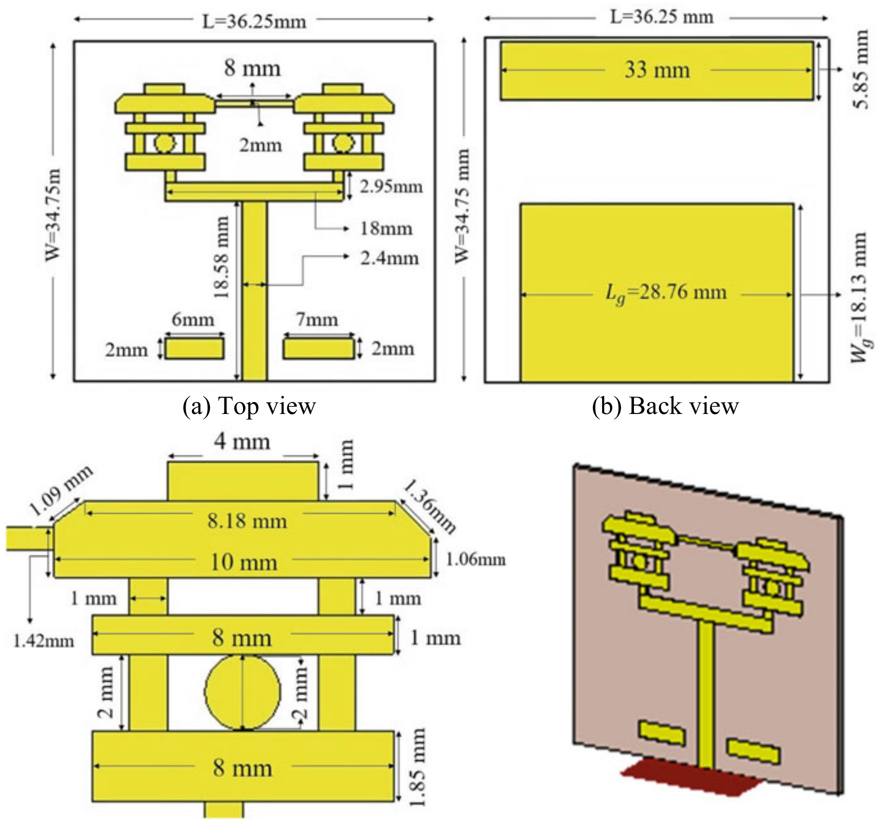
2 Design and Structure of the Antenna

The microstrip patch array antenna consists of two patches as appeared in Fig. 1a mounted on Roger RT 5880 (lossy) substrate having a compact size of 36.25 × 34.75 mm^2. Substrate thickness plays a significant role in performance of the design. The thickness of the selected Roger RT 5880 is 0.79 mm as well as the thickness (copper) of the patch elements and modified ground plane including parasitic elements is 0.035 mm.

The length and width of the feed are 18.58 and 2.4 mm, respectively. Three parasitic elements are used. Two are on the dielectric substrate, and the other is on the ground plane. The length and width of the parasitic elements placed on the dielectric substrate are 6 mm, 2 mm and 7 mm, 2 mm, respectively. The other parasitic element which is placed on the ground plane has a length of 33 mm and a width of 5.85 mm as illustrated in Fig. 1b. The thickness of these parasitic elements is also 0.035 mm and is made of Copper. Two connectors are carried out from the patches. The connector which is placed in the upper portion between the patches has 8 mm of length and 2 mm of width. And the other connector taken from the patches and is connected to the feed has a length of 18 mm and a width of 2.95 mm. The amplified view in Fig. 1c depicts the single element of the patch of the UWB CMPAA. Figure 1d illustrates the 3D view of the compact antenna. The important factors of the CMPAA are enlisted in Table 1.

3 Simulation, Result and Discussion

The compact UWB MPAA is designed and optimized by using CST-MWS. The simulated scattering parameter (S_{11} parameter) of the patch array antenna is presented



(a) Top view (b) Back view

(c) Amplified view (d) 3D view

Fig. 1 Compact UWB MPAA

in Fig. 2. There are four lobes within the entire frequency band. The reflection coefficients are −32.8, −35.33, −22.524, and −18.23 dB at the frequencies of 3.45, 4.68, 7.78, and 12.096 GHz, respectively. The −10 dB impedance bandwidth is 10.08 GHz (2.77 GHz to 12.85 GHz) which covers the UWB spectrum frequency as determined by FCC. The 3D gain at the resonant point of 4.68 GHz, gain (Cartesian) at 4.68 GHz and 3D directivity at 4.68 GHz of the patch array antenna is illustrated in Fig. 3a, b and c and afterward, the linear gain and directivity graph is plotted in Fig. 3d.

From the illustrated figures, it is seen that at 4.68 GHz, the gain and the directivity are 3.92 dB and 4.55 dBi, respectively. However, since the output curve covers a huge frequency band from 2.77 to 12.85 GHz, it has four lobes in total, and we observed gain and directivity at all the points within the coverage band which reflects the antenna achieves good gain and directivity within the entire spectrum frequency. The 2.86, 5.79, and 5.97 dB are the gains at 3.45, 7.8, and 12.09 GHz, respectively,

Table 1 Factors of the designed compact UWB MPAA

Description and symbol	Value (mm)
Substrate's length (L)	36.25
Substrate's width (W)	34.75
Substrate's thickness (h)	0.79
Metal's thickness (t)	0.035
Length of the parasitic-1	6
Width of the parasitic-1	2
Length of the parasitic-2	7
Width of the parasitic-2	2
Length of the parasitic-3	33
Width of the parasitic-3	5.85
Ground plane's length (L_g)	28.76
Ground plane's width (W_g)	18.13
Feed's length	18.58
Feed's width	2.40
Radius of circle	1
Stub's length	4
Stub's width	1

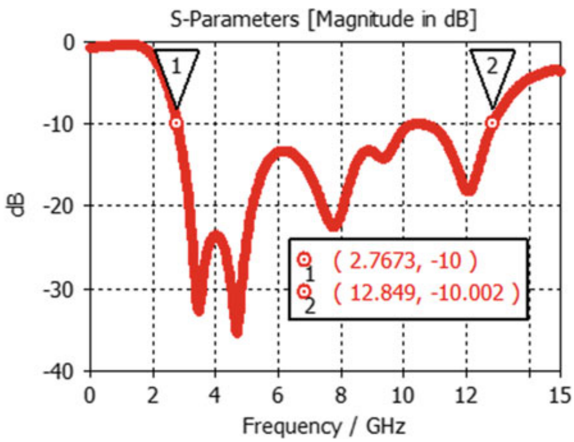
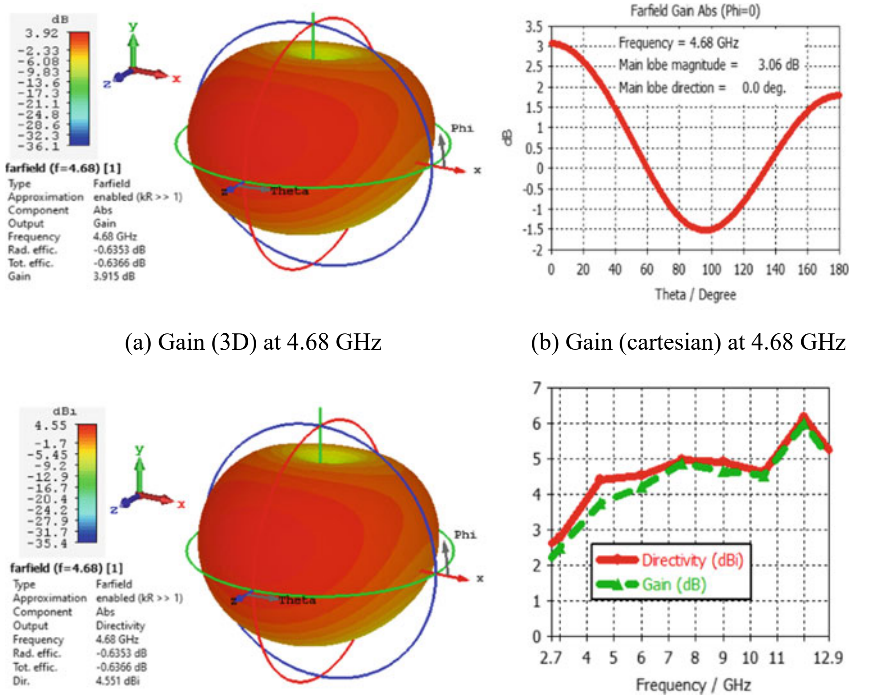


Fig. 2 S_{11} curve of the designed compact UWB MPAA

of the remaining three lobes. Similarly, 3.08 dBi, 5.95 dBi, and 6.13 dBi are the directivities at 3.45 GHz, 7.8 GHz, and 12.09 GHz, respectively. It is also observed from Fig. 3d that the gain and directivity vary from 2.28 to 5.97 dB and 2.65 dBi to 6.14 dBi, respectively, within the large operating frequency. Hence, the maximum gain is 5.97 dB and maximum directivity is 6.14 dBi which makes the CMPAA suitable for UWB spectrum. The polar plots representation of the radiation patterns (both E-fields and H-fields) of the compact UWB MPAA for $\phi = 0°$ and 90° at



(a) Gain (3D) at 4.68 GHz

(b) Gain (cartesian) at 4.68 GHz

(c) Directivity (3D) at 4.68 GHz

(d) Gain and directivity vs. frequency

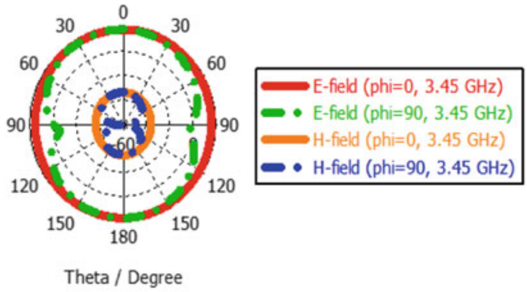
Fig. 3 Gain and directivity curve of the designed compact UWB MPAA

3.45, 4.68, 7.8, and 12.09 GHz are plotted in Fig. 4. For 3.45 GHz, the maximum half power beam width (HPBW) is 83.7 ° at φ = 90°. At φ = 0°, the prime lobe magnitude is 17.5 dBV/m for E-field and −34 dBA/m for H-field. Similarly at φ = 90°, the prime lobe magnitude is 17.6 dBV/m for E-field and −33.9 dBA/m for H-field. The prime lobe is directed at 0 ° for φ = 0° while the main lobe is directed at 8 ° for φ = 90°. For the remaining frequencies of 4.68, 7.8 and 12.09 GHz, the polar plot values of the radiation patterns for E-fields and H-fields are demonstrated in Tables 2 and 3, respectively.

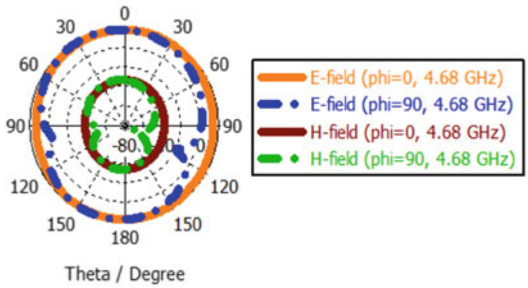
The VSWR of the compact UWB MPAA is shown in Fig. 5. The value of VSWR fluctuates within 1−2 throughout the entire frequency band, and it is 1.0348 at the resonant point of 4.68 GHz which implies good port impedance matching characteristics of the compact UWB antenna. In Fig. 6, antenna efficiency is plotted with respect to the operating spectrum. Radiation efficiency can be defined as the ratio of gain to directivity. The maximum antenna efficiency is 98% and the minimum efficiency is 86.4% at 4.68 GHz which reflects that the designed compact UWB array antenna radiates most of the received power.

In Fig. 7, the surface current distribution is displayed at four different frequencies. The maximum surface currents are 146.457 A/m at 3.45 GHz, 145.613 A/m

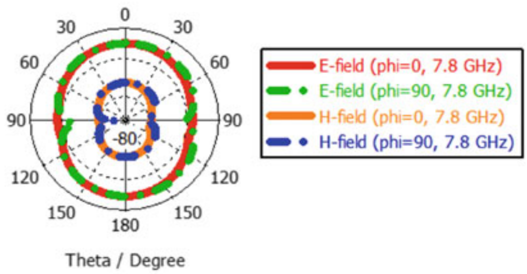
Fig. 4 E-field and H-field curve of the proposed compact UWB MPAA



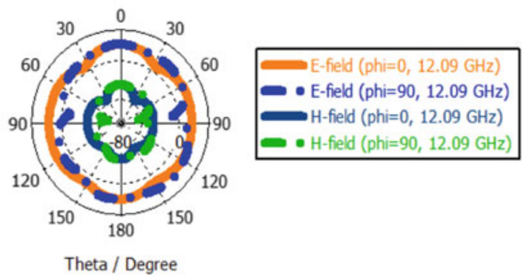
(a) 3.45 GHz



(b) 4.68 GHz



(c) 7.8 GHz



(d) 12.09 GHz

Table 2 E-field properties

Index	E-field					
	4.68 GHz		7.8 GHz		12.09 GHz	
	φ = 0°	φ = 90°	φ = 0°	φ = 90°	φ = 0°	φ = 90°
Lowest side lobe level (LSLL) (dB)	−1.3	−1.7	0	0	−2.2	−1.2
Half power beam width (HPBW)	113.8 °	74.3 °	62.9 °	81.2 °	31.4 °	67.9 °
Main lobe Magnitude (dV/m)	17.8	18.7	20.4	20.5	20.7	20.7
Main lobe direction	0 °	20 °	180 °	169 °	0 °	1 °

Table 3 H-field properties

Index	H-field					
	4.68 GHz		7.8 GHz		12.09 GHz	
	φ = 0°	φ = 90°	φ = 0°	φ = 90°	φ = 0°	φ = 90°
Lowest side lobe level (LSLL) (dB)	−1.3	−1.7	0	0	−2.2	−1.2
Half power beam width (HPBW)	113.8 °	74.3 °	62.9 °	81.2 °	31.4 °	67.9 °
Main lobe magnitude (dV/m)	−33.7	−32.8	−31.2	−31	−30.8	−30.8
Main lobe direction	0 °	20 °	180 °	169 °	0 °	1 °

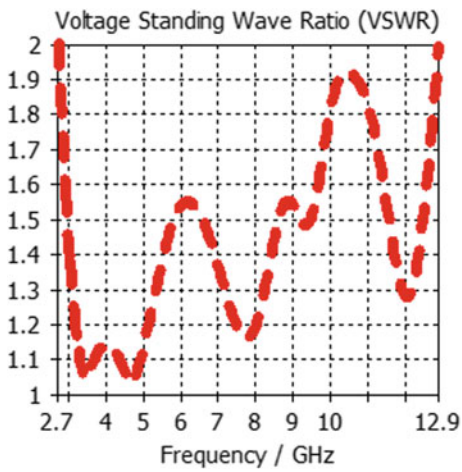


Fig. 5 VSWR

at 4.68 GHz, 161.538 A/m at 7.8 GHz, and 165.481 A/m at 12.09 GHz. The Z-parameters of the UWB CMPAA is illustrated in Fig. 8 which indicates that the real part is close to 50 Ω (48.9 Ω), while the imaginary value is close to 0 Ω (−1.6266 Ω) at 4.68 GHz.

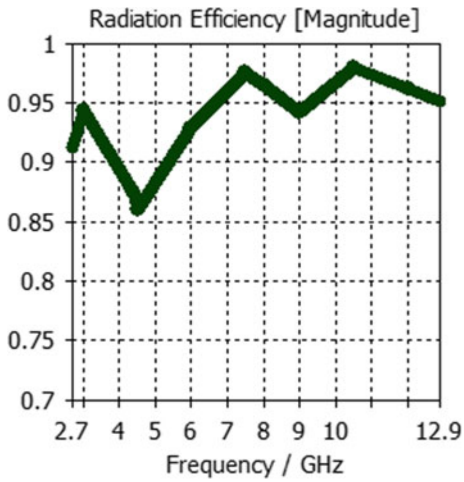


Fig. 6 Radiation efficiency

The impedance matching is very good for the antenna with respect to 50Ω. The overall antenna output results are summarized in Table 4. From Table 5, our proposed antenna possesses compact size with maximum bandwidth and maximum gain. In [4] and [14], though the UWB antennas have slightly lower volume than our proposed UWB antenna, the presented UWB antenna in this paper has very larger bandwidth as well as better gain.

4 Conclusion

The proposed compact microstrip patch array antenna (995.153 mm^3) with a thickness of 0.79 mm having a modified ground plane and parasitic elements generates an operating frequency band ranging from 2.77 to 12.85 GHz. The main objective of the antenna was to obtain ultra wide operating band and compactness, and the design array antenna entirely covers UWB spectrum (3.1−10.6 GHz) as well as secures its compact size without compromising other performance parameters like gain, VSWR, and efficiency. The impedance requirements are also obtained as expected. The maximum gain (5.97 dB) and maximum radiation efficiency (98%) of the compact UWB MPAA are also found very satisfactory. The optimized parameters used in this design provide the effective outcome for which the designed antenna is capable of working appropriately in the target frequency bands, and the achieved bandwidth meets the UWB requirement and the FCC's standardization. As a result, the proposed compact UWB patch array antenna can be used for microwave imaging (3.1–10.6 GHz), WiMAX (3.4–3.6 GHz), WLAN, WiFi-5/6 (5.180−5.925, 5.940–7.075 GHz), Sub-6 GHz (3.3–4.9 GHz), and many other wireless applications.

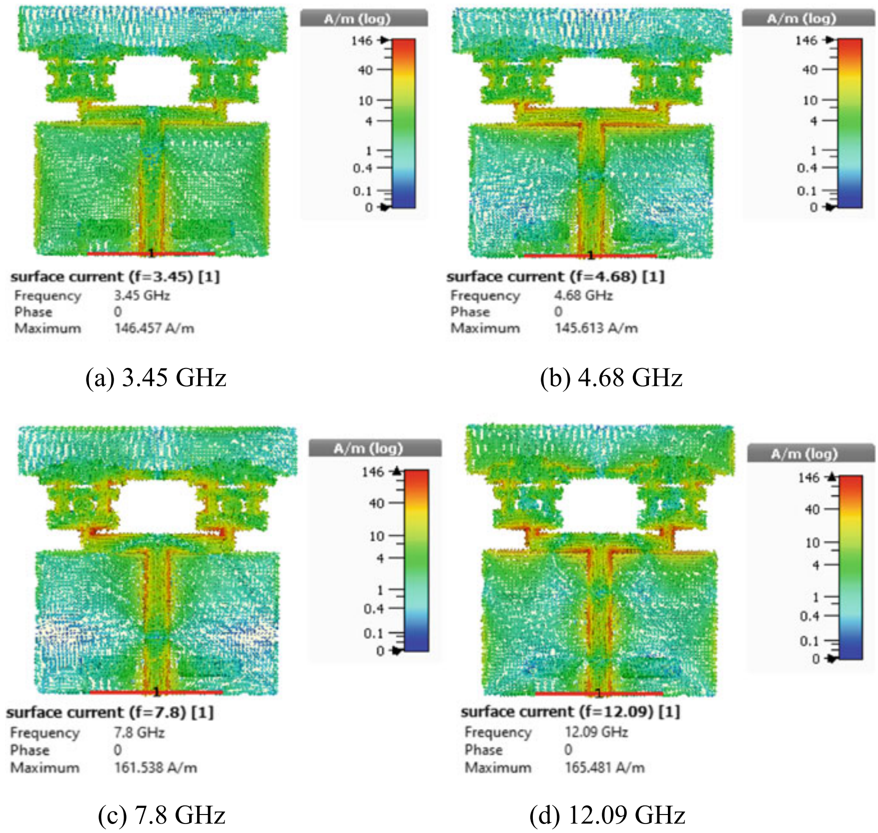


Fig. 7 Surface current at different frequencies

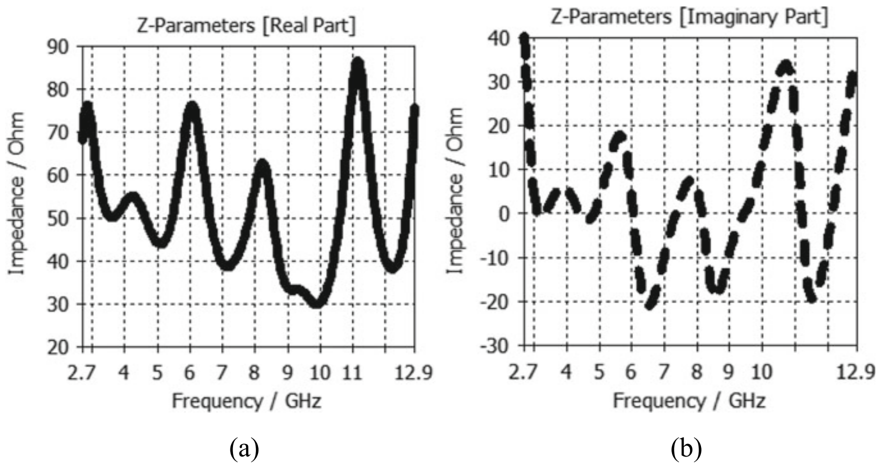


Fig. 8 Z-parameters **a** Real part and **b** Imaginary part

Table 4 Results of the designed compact UWB MPAA

Description	Value (mm)
Lower cut-off	2.77 GHz
Upper cut-off	12.85 GHz
−10 dB Bandwidth	10.08 GHz
Resonant frequency	4.68 GHz
Reflection coefficient at 4.68 GHz	−35.33 dB
VSWR at 4.68 GHz	1.0348
Maximum gain	5.97 dB
Maximum directivity	6.14 dBi
Maximum radiation efficiency	98%
Impedance (Real Part)	48.9 Ω
Impedance (Imaginary Part)	−1.6266 Ω

Table 5 Comparison table

Index	Reference number				This work
	[4]	[14]	[15]	[16]	
Year	2017	2019	2017	2019	2021
Size (L × W × h) mm^3	921.6	858.624	1440	2032	995.15
Substrate thickness (mm)	1.6	1.6	1.6	0.508	0.79
Substrate material (dielectric constant, tangent loss)	FR-4 (4.4, 0.0019)	FR-4 (4.4, 0.02)	FR-4 (4.4, 0.02)	Teflon (2.55, 0.002)	Rogers RT 5880 (2.2, 0.0009)
Operating frequency range (GHz)	3.4–9.4	5.86–6.37	2.9–12.4	2.32–5.24	2.77–12.85
No. of prime resonant point (GHz)	2	1	4	2	4
Resonant point (GHz)	4.4, 8	6.08	1.5, 3.5, 5.2, 5.8	2.55, 4.31	3.45, 4.68, 7.8, 12.09
Reflection coefficient (dB)	≈−31.5, ≈−22	−18.58	≈ −44	≈ 24.9, ≈−22.5	−35.33
Maximum gain (dB)	≈ 2.5	2.98	3.01	4.31	5.97
−10 dB BW (GHz)	6	0.503	9.5	2.92	10.08Fi

References

1. Singh, C., Kumawat, G.: A compact rectangular ultra-wideband microstrip patch antenna with double band notch feature at Wi-Max and WLAN. Wireless Pers. Communi. **114**, 2063–2077 (2020). https://doi.org/10.1007/s11277-020-07465-1
2. Saeidi, T., Ismail, I., Wen, W.P., Alhawari, A.R., Mohammadi, A.: Ultra-wideband antennas for wireless communication applications. Int'l J. Antennas and Propagation. **2019**, 1–25 (2019). https://doi.org/10.1155/2019/7918765
3. Malekpoor, H., Hamidkhani, M.: Compact multi-band stacked circular patch antenna for wideband applications with enhanced gain. Electromagnetics **39**(4), 241–253 (2019). https://doi.org/10.1080/02726343.2019.1595379
4. Sharma, A., Khanna, P., Kumar, A., Shinghal, K.: A CPW-fed structure shaped substrate wideband microstrip antenna for wireless applications. J. Microwav. Optoelectron. Electromagnet. Appl. **16**, 419–433 (2017). https://doi.org/10.1590/2179-10742017v16i2805
5. Paul, L.C. et al.: A novel miniaturized coplanar waveguide fed tapered slot ultra wide band Vivaldi antenna for microwave imaging applications. In: 10th Int'l Conf. on Computing, Communication and Networking Technologies. pp. 1–6. (2019). https://doi.org/10.1109/ICCCNT45670.2019.8944602
6. Khandelwal, M.K.: Analysis of frequency notch characteristics in enhanced-ultra-wide-band compact microstrip patch antenna. Wireless Pers. Commun. **100**, 1249–1260 (2018). https://doi.org/10.1007/s11277-018-5638-0
7. Paul, L.C. et al.: Human brain tumor detection using CPW Fed UWB Vivaldi antenna. In: IEEE Int'l Conference on Biomedical Engineering, Computer and Information Technology for Health. pp. 1–6. (2019). https://doi.org/10.1109/BECITHCON48839.2019.9063185
8. Hamad, E.K., Mahmoud, N.: Compact tri-band notched characteristics UWB antenna for WiMA WiMAX, WLAN and X-band applications. Adv. Electromagn. **6**(2), 53–58 (2017). https://doi.org/10.7716/aem.v6i2.465
9. Abedin, Z.U., Ullah, Z.: Design of a microstrip patch antenna with high bandwidth and high gain for UWB and different wireless applications. Int'l J. Adv. Comput. Sci. Appl. **8**(10), 379–382 (2017). https://doi.org/10.14569/IJACSA.2017.081049
10. Al-Faouri, Y.S., Awad, N.M., Abdelazeez, M.K.: Enhanced ultra-wide band hexagonal patch antenna. Jordanian J. Comput. Info. Techno. **4**(3), 150–158 (2018). https://doi.org/10.5455/jjcit.71-1534239855
11. Gupta, M., Mathur, V.: Koch boundary on the square patch microstrip antenna for ultra wideband applications. Alexandria Eng. J. **57**(3), 2113–2122 (2018). https://doi.org/10.1016/j.aej.2017.06.005
12. Rahman, N., Islam, M.T., Mahmud, Z., Samsuzzaman, M.: The broken-heart printed antenna for ultrawideband applications. design and characteristics analysis. IEEE Ant. Propaga. Magazine, **60**(6), 45–51(2018). https://doi.org/10.1109/MAP.2018.2870664
13. Messaoudene, I., Denidni, T., Benghalia, A.: Low-profile U-shaped DRA for ultra-wideband applications. Int. J. Microw. Wirel. Technol. **9**(3), 621–627 (2017). https://doi.org/10.1017/S1759078716000155
14. Othman, A., Shaari, N.I.S., Zobilah, A.M., Shairi, N.A., Zakaria, Z.: Design of compact ultra wideband antenna for microwave medical imaging application. Indonesian J. Electri. Eng. Comput. Sci. **15**(3), 1197–1202 (2019). https://doi.org/10.11591/ijeecs.v15.i3.pp1197-1202
15. Ali, T., Biradar, R.C.: A miniaturized Volkswagen logo UWB antenna with slotted ground structure and metamaterial for GPS, WiMAX and WLAN applications. Progress In Electromagn. Res. C **72**, 29–41 (2017). https://doi.org/10.2528/PIERC16120109
16. Tang, X et al.: Ultra-wideband patch antenna for sub-6 GHz 5G communications. In: Int'l Workshop on Electromagnetics: Applications and Student Innovation Competition. pp. 1–3. (2019). https://doi.org/10.1109/iWEM.2019.8887933

An Integrated Framework for Bus-Driver Allocation in Bangladesh

Mondira Chakraborty and Sajeeb Saha

Abstract Intelligent transport system (ITS) aims to launch constructive vehicular connectivity and synchronization techniques which notably improve travel efficiency by reducing traffic gridlock, irregularities, mishaps, etc. There are numerous applications of ITS innovations in first world countries where emergent nations are yet depending on the fragmented disorganized analog public transportation system. The key challenge in underdeveloped countries is the maintenance of efficient timetables and scheduling of public transportation systems. In this paper, we have proposed a bus transit framework that consists of a timetable and a bus-driver scheduling algorithm to resolve this predicament. A coherent timetable considering passenger load in traffic rush hours is generated for scheduling the fixed number of trips and lines. To schedule the buses and drivers, three selection procedures—First Serve Check First (FSCF), Equal Allocation (EA), Random Allocation (RA) have been implemented and analyzed with varying numbers of trips and lines. The experimental results depict that, FSCF schedules with a minimum number of vehicles and drivers, though the algorithm runtime is higher compared to other approaches.

Keywords Public transport framework · Intelligent transportation system (ITS) · Bus-driver scheduling

1 Introduction

Intelligent transport system (ITS) is a quantum leap in sustainable transportation technology as it aims to innovative services relating to different modes of transport and traffic management. At present, well-developed countries are using different neoteric forms of ITS to reduce crimes, traffic violations, and to enhance security. Some major applications of ITS are vehicle management, smart traffic management, speed alerts, RFID in freight transportation, variable speed limits, dynamic traffic light sequence, collision avoidance systems, etc. A dynamic vehicle routing

M. Chakraborty · S. Saha (✉)
Jagannath University, Dhaka, Bangladesh
e-mail: sajeeb@cse.jnu.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_26

and scheduling model that incorporates real-time information using variable travel times is a necessity for any smart city. On the contrary, in third world countries, underdeveloped and impoverished transportation systems combined with an enormous population make it critical to organize and maintain trip schedules and main resources manually at an operational level.

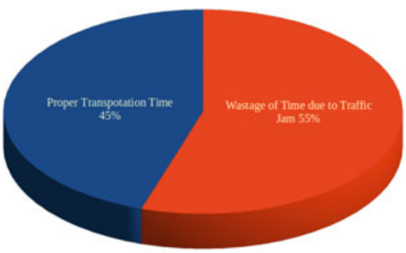
As an example, our case study shows that Dhaka, the capital of Bangladesh which is an inhabitant of about 20.2 million people in total 306.38 km^2 area, needs nearly 30 million trips per day [1, 2]. Cater to these huge schedules, the bus is the only major available public transport system, as buses carry 47% of the total trip and 70% of motorized trip [2]. Handling this rapid growth of population, the number of public transports like buses is also increasing in a disorganized, ill-disciplined manner which is causing more traffic violations and traffic gridlocks. Every year more than 1000 buses and minibuses are registered in Dhaka [2]. According to the registration records of BRTA, the number of buses and minibuses registered in Dhaka from 2009 until 2016 is available in Table 1.

In this megacity, people on average spend 2.35 h in the traffic of which 1.30 h are due to traffic jams which means wastage of 1.30 man-hour of resource per day [3]. Fig. 1 shows that people are loosing 55% of their traffic hours during staying in the traffic [3]. Similarly, vehicle drivers are also loosing almost three hours due to traffic jams every day. On average, their daily working hour is slightly over 12 h. It means they are losing almost 25% of their working hour [3].

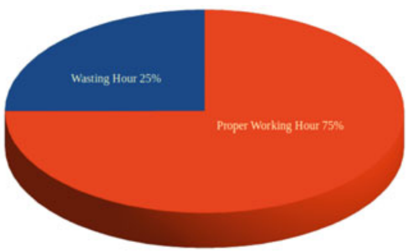
In the context of sustainable urban development, next-generation public transportation (PT) services along with ITS are receiving considerable attention. Over

Table 1 Number of registered buses and minibuses in Dhaka (yearly) [2]

Year	2009	2010	2011	2012	2013	2014	2015	2016	Grand total
Bus	914	1231	1501	1218	971	1364	2221	3419	27,537
Minibus	112	149	136	103	83	135	103	164	10,214
Total	1026	1380	1637	1321	1054	1499	2324	3583	37,751



(a) General people



(b) Vehicle drivers

Fig. 1 Wastage of time due to traffic jam [3] the years, many researchers have proposed different approaches to deal with problems related to the bus-driver scheduling problem such as vehicle scheduling problem (VSP) and crew rostering (CR) problem. Few state-of-the-art used column generation techniques combined with the genetic algorithm [4]. Besides these approaches, some studies used heuristic approaches like hyper-heuristic [5], meta-heuristic [6], variable neighborhood search heuristic [6], hybrid heuristic combined with greedy randomized adaptive search procedure (GRASP) [7], etc. There are some other approaches, namely fuzzy genetic [8], unicost set covering problem (SCP) combined with column generation [9] which are used to solve the bus-driver scheduling problem as well. An appropriate and optimal scheduling procedure is a necessity to reduce traffic violations, a high rate of roadside mishaps, traffic crimes, and gridlocks. The key contributions of this work are summarized as follows:

- We Design a bus-driver scheduling framework considering the driver workload and rush-hour traffic demand.
- We develop an allocation algorithm with three strategies to schedule the trips with the minimum number of buses and drivers.
- We conduct an extensive evaluation to analyze the performance of the proposed strategies.

The rest of the paper is organized as follows: Sect. 2 contains a comprehensive literature survey on transport scheduling problems and possible solutions, Sect. 3 explains the significance and impact of the proposed approach and procedure of implementing the framework, and Sect. 4 outlines the comparative analysis in terms of different performance indicators. Finally, Sect. 5 summarizes the result of our analysis and the future scope of our work.

2 Literature Review

In this paper, a bus-driver scheduling timetable which is similar to various generic problems like vehicle scheduling problem (VSP), crew assignment problem (CAP), or crew rostering (CR) is designed to fit perfectly with the various customized necessities of authority. There are two related mathematical solutions which are the classical set covering problem (SCP) and the set partition problem (SPP). Both of these problems are NP-hard problems, and different techniques to handle these problems are immensely outlined in the Operational Research Literature (ORL) [10]. A small number of papers have discussed the use of a composite version of column generation (CG) with other procedures, e.g., [11] to solve instances which can be up to 138 tasks, one paper analyzed the use of CG combined with a genetic algorithm (GA) [5].

There are two types of integrated approaches:(i) partial integration and sequential solution procedures and (ii) complete integration where decisions of the integrated problems must be taken simultaneously [12]. Some completely integrated approaches assume the existence of a homogeneous fleet and generic drivers. This assumption implies that the vehicles have the same characteristics and can be assigned to the same set of trips, and drivers are alike and can be assigned to any feasible duty without considering their qualifications. If heterogeneous fleets are considered, the vehicle scheduling problem (VSP) must identify the type of vehicle required to cover each trip. However, vehicles are commonly considered alike within each type of fleet. This problem is a particular case of the multi-depot vehicle scheduling problem (MDVSP) where one depot has a single fleet type and it can be formulated as a network flow problem. The MDVSP has been integrated with the CSP [13]. Another heuristic method called hybrid heuristic that combines greedy randomized adaptive search procedure (GRASP) and Rollout meta-heuristics can solve instances of up to 250 tasks. An algorithm based on iterated local search (ILS) similar to variable neighborhood search and very large-scale neighborhood search (VLNS), as well as a self-adjusting algorithm, based on evolutionary approach is also used to tackle the problem.

3 Methodology

Our approach is moderately focused on the geographical area like Dhaka city, which includes circumstances of huge population density, regular traffic congestion, fragmented public transport system with low or no usage of ITS. Our proposed method is delineated to inaugurate a proper and defined bus-driver scheduling model which can be customized and modified according to a specific environment's needs. Dhaka city's total number of trips is nearly 20.5 million where transportation modal share is approximately car—5.1%, rickshaw—38.3%, walk—19.8%, public bus—28.3%, private bus—1.8%, CNG—6.6%, railway—0%, and lastly, the total number of bus routes or line in Dhaka city is 149 [14]. Each depot considers a group of lines or routes. In this approach, our goal is to schedule one depot which consists of a few lines or routes. The scheduling process is composed of four major distinct events which are described below:

1. **Line Planning**: The concept of line relies on the idea of a route with a fixed round time to complete the trip and return to the depot. This prototype combines choosing the option of more than one line, and each line contains a schedule of trips. Let $L = \{l_1, l_2, l_3, ..., l_m\}$ where $l_i \in L$, which means set of lines.
2. **Trip Scheduling**: The trip schedule consists of a predetermined time range where after a discrete-time gap a set of trips is allocated, it is decided according to peak-hour, off-hour, and each line's passenger load. Let $T = \{t_1, t_2, t_3, ..., t_k\}$ where $t_i \in T$, which is a specific time range and time gap is $\alpha = t_d - t_{d-1}$ as $1 < d \leq k$.
3. **Driver Scheduling**: Each trip is assigned with a driver by maintaining all the labor and transportation laws. Drivers are elected in a way such that they do not collide with more than one trip at a time and their working hours decently gap. Let $D = \{d_1, d_2, d_3, ..., d_n\}$ where $d_i \in D$, which means set of drivers.
4. **Bus Scheduling**: In bus scheduling, every trip is also allocated with a bus to cover a trip. Fit and appropriate buses are selected and paired up with drivers to

complete the trips. In this selection procedure, it is also checked if any vehicle is working continuously for 4 h without any break or any vehicle crossed the 8-h limit. Let $V = \{v_1, v_2, v_3, ..., v_n\}$ where $v_i \in V$ which means set of buses.

To create a better visual representation, let us assume that a bus agency of Dhaka city discharges a preset number of trips from one depot with two precise lines which has a complete trip round time of about one hour for Line 1 and two hours for Line 2 per day. It is also designed concerning the passenger load of peak and off-hour time. Here, drivers and vehicles need to be assigned in the most optimal way possible so that each driver gets a break after the continuous 4h journey and work limit of 8 hour. Our main objective is to reduce the number of drivers and vehicles needed to cover the total trip count.

Table 2 describes a realistic version of the total timetable and scheduling result for two distinct lines which shows the arrangements of a part-time range. It shows the trips reserved for a time gap of half-hour grouped with specific drivers and vehicles. As found in our case study, 7:00 AM to 10:00 AM is a rush-hour time span, and clearly, the passenger load is heavy during this period, so more general public transportation like the bus is required to meet the expectations. The percentage of total trip allocation during this peak and off-hour, total trip time, number of lines, trip distribution for each line, etc. can be set by authorized customization and particular region's demand. According to our case study, we have set 60% of total trips during peak hours and the rest 40% to off-hour. We also set the number of trips, drivers, and vehicles equal for drafting our model.

Our proposed model's algorithm is divided into two portions: Timetable Dataset Generation and diver-vehicle mapping according to generated data. Timetable Dataset Generation creates a timetable for each line according to the basic Dhaka city trip assignment, and the mapping algorithm maps drivers and vehicles to the listed trips. The major notations used in the algorithms are listed in Table 3.

3.1 Timetable Dataset Generation

To prepare the trip schedule, an algorithm is developed that creates a dataset of the trips based on the trip time, the number of lines and sets the interval between two trips according to real scenarios of the Dhaka bus route (Lines 1–2). In our general study, it is found that 7:00 am to 10:00 am and 5:00 pm to 8:00 pm are considered as rush hours in Dhaka city. So, the allocation of trips for each line's schedule according to peak-hour and off-hour time period is decided (Lines 3–9). In the state-of-the-art works, the maximum number of trips considered 50 per line with 10 fixed intervals [15]. In this paper, a half-hour time gap has been considered though it can be customized according to the system requirement. The round time for each trip is different in every line ranging from 1 to 4 h. The number of lines ranges from 2 to 19 and trip number from 80,000 to 200,000, which is relatively higher in every spectrum of the state-of-the-art sample data. These generated datasets are used to evaluate the performance of the studied algorithms.

Table 2 Solution for one depot scheduling

Time	Line 1			Line 2			Time	Line 1			Line 2		
	Trip	Driver	Vehicle	Trip	Driver	Vehicle		Trip	Driver	Vehicle	Trip	Driver	Vehicle
6:00	1	d-100	v-100	1	d-101	v-101	9:00	3	d-100	v-100	3	d-106	v-106
6:30	1	d-102	v-102	1	d-103	v-103			d-101	v-101		d-107	v-107
7:00	3	d-100	v-100	3	d-106	v-106			d-104	v-104		d-108	v-108
		d-104	v-104		d-107	v-107	9:30	2	d-102	v-102	2	d-109	v-109
		d-105	v-105		d-108	v-108			d-103	v-109		d-111	v-111
7:30	3	d-102	v-102	3	d-111	v-111	10:00	2	d-100	v-100	2	d-104	v-104
		d-109	v-109		d-112	v-112			d-101	v-101		d-105	v-105
		d-110	v-110		d-113	v-113	10:30	1	d-102	v-102	1	d-103	v-103
8:00	3	d-100	v-100	3	d-105	v-105	11:00	1	d-100	v-100	1	d-101	v-101
		d-101	v-101		d-114	v-114	11:30	1	d-102	v-102	1	d-106	v-106
		d-104	v-104		d-115	v-115	12:00	1	d-100	v-100	1	d-104	v-104
8:30	3	d-102	v-102	3	d-110	v-110	12:30	1	d-102	v-102	1	d-103	v-103
		d-103	v-103		d-116	v-116	1:00	1	d-100	v-100	1	d-101	v-101
		d-109	v-109		d-117	v-117	1:30	1	d-102	v-102	1	d-105	v-105
							2:00	1	d-104	v-104	1	d-106	v-106

Table 3 List of notations

Symbol	Description
$\mathcal{N}$	Total trip number
$\mathcal{T}$	Total scheduling period
α	Time gap (explain in trip scheduling part)
$\mathcal{L}$	Line's list
$\mathcal{D}$	Driver's list
δ	Drivers in break list
ϵ	Drivers in instant available list
ω	Drivers in working list
$\mathcal{V}$	Vehicle's list
ν	Vehicle's in service list
θ	Vehicles in instant available list
η	Vehicles in working list
ρ	Peak-hour trip demand
σ	Off-hour trip demand
b_l^t	Number of assigned trips in $t \in \mathcal{T}$ time of $l \in \mathcal{L}$ line
$\varDelta_d$	Working hour of $d \in \mathcal{D}$
$\varDelta_v$	Working hour of $v \in \mathcal{V}$
Γ_d	Trip completion time of $d \in \mathcal{D}$
Γ_v	Trip completion time of $v \in \mathcal{V}$
$\varLambda_d$	Continuous working hour of $d \in \mathcal{D}$
$\varLambda_v$	Continuous working hour of $v \in \mathcal{V}$
$\varOmega_d$	Total working hour of $d \in \mathcal{D}$
$\varOmega_v$	Total working hour of $v \in \mathcal{V}$

Algorithm 1: Timetable Generation()

1. Initialization
2. Set $\mathcal{N}$, $\mathcal{L}$, $\mathcal{T}$, α, ρ and σ
3. **for** $\mathcal{L} = \{l_1, l_2, l_3, ..., l_m\}$ **do**
4.   **for** $\mathcal{T} = \{t_1, t_2, t_3, ..., t_k\}$ **do**
5.   Decide t_j as peak time or off time
6.   Allocate b_l^t according to ρ or σ
7.   **end for**
8. **end for**
9. Print the Scheduling

3.2 Driver-Vehicle Mapping

Algorithm (2) dvvhcMapping() allocates a pair of drivers and buses to each trip so that drivers or vehicles do not get assigned in more than one trip at the same time. It starts by initializing variables (Line 1). For each distinct time gap, the driver's list

Algorithm 2: dvvhcMapping()

INPUT: *$\mathcal{N}$, $\mathcal{D}$, $\mathcal{V}$, each line's timetable*
OUTPUT: *Mapping between driver and vehicle for each line's timetable*

1. Initialization
2. **for** $\mathcal{T} = t_1, t_2, t_3, ..., t_k$ **do**
3.   Update δ and ν
4.   **if** $\epsilon \neq 0$ **then**
5.   **for** $L = \{l_1, l_2, l_3, ..., l_m\}$ **do**
6.   **while** $b_l^t \neq 0$ **do**
7. $trp \in b_l^t$
8. $b_l^t \leftarrow b_l^t / trp$
9.   Select drv from ϵ and vhc from θ
10.   Mark (trp, drv, vhc) as assigned and not instant available
11.   **end while**
12.   **end for**
13.   **end if**
14.   **if** $\omega \neq 0$ **then**
15.   Update Δ_d and Δ_v where $d \in \omega$ and $v \in \eta$
16.   **if** $\Delta_d == \Gamma_d$ **then**
17.   Mark $d \in \omega$ on break if $\Lambda_d \geq 4hr$
18.   Mark $d \in \omega$ as not available if $\Omega_d \geq 8hr$
19.   Otherwise, mark $d \in \omega$ as instantly available
20.   Remove d from ω list
21.   **end if**
22.   **if** $\Delta_v == \Gamma_v$ **then**
23.   Mark $v \in \eta$ on break if $\Lambda_v \geq 4hr$
24.   Mark $v \in \eta$ as not available if $\Omega_v \geq 8hr$
25.   Otherwise, mark $v \in \eta$ as instantly available
26.   Remove v from η list
27.   **end if**
28.   **end if**
29. **end for**

in the short break and the vehicles in the service list get updated; if they completed the break time, they are made available for the next trip assignment (Lines 2–3).

For each line, the driver and vehicle for each trip are allocated from the available list for a particular time. Three algorithms, namely Random Allocation (RA), Equal Allocation (EA), First Serve Check First (FSCF), are implemented separately to select drivers and vehicles. RA selects drivers and buses spontaneously without any precondition. EA provides a chance to every driver and vehicle to complete a trip, and after that, if any trips are left, drivers and vehicles are repeated. FSCF chooses drivers or vehicles from the available driver or vehicle list and always starts selecting from the top of the list. This procedure repeats any drivers and vehicles who are again available, after completing a trip. Usually, this algorithm focuses on utilizing drivers and vehicles most optimally. Then after selection, those units are marked as not available until the trip completes (Lines 4–13). After that, the drivers in the working list, which are assigned to a trip, are updated and checked if any of those completed the assigned running trip. If any drivers are continuing trips for 4 h or more, then the

Table 4 Performance evaluation metrics

Metrics	Description	Range
Trips	Number of trips covered by all lines	80,000–200,000
Lines	Number of lines with a fixed round timenewline to complete per trip	2–12
Round time	Time to complete one trip and return	1–4h
Runtime	Algorithm's execution time	In millisecond

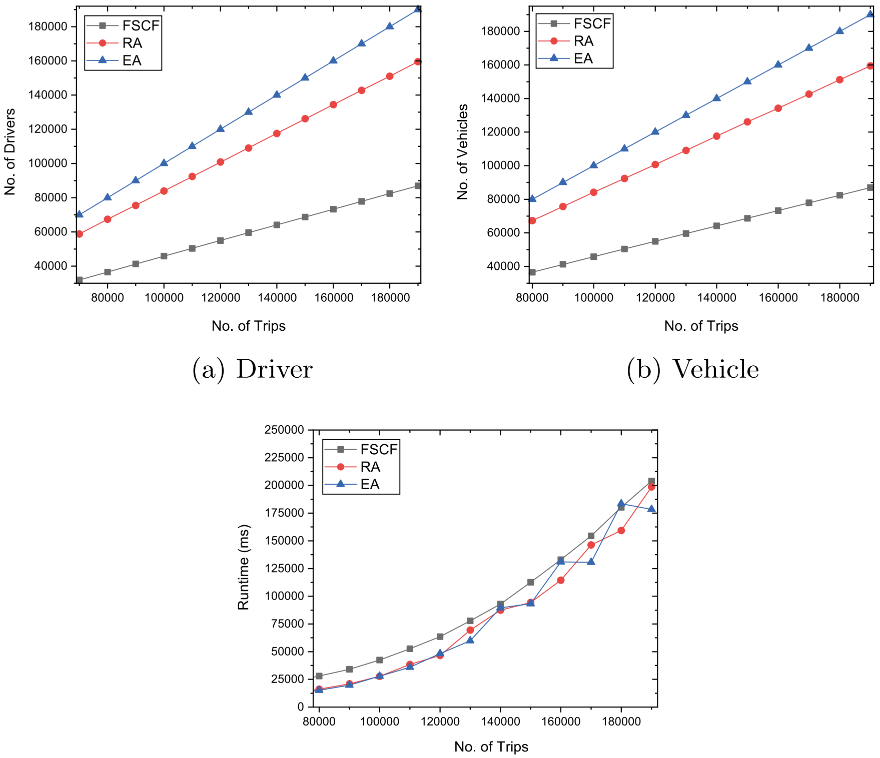
drivers are given half hours break. If drivers complete 8 working hours excluding break hours, then those drivers are restricted from any more trips (Line 19). If any driver engaged in longer trips that cross the break time constraint, then a break is assigned immediately after completing the trip (Lines 14–21).

After completion of a trip, a vehicle is checked if it is running continuously for 4 h then it is sent to servicing or if it is working for 8 h, then it is restricted for any more trips. If none of the conditions work, then it is made available for another trip. Eventually, removing the vehicle from the working list and thus repeating this whole procedure is followed for each time gap (Lines 22–29).

4 Performance Evaluation

In this section, we evaluate the performance of the proposed bus-driver allocation strategies. C++ programming language is applied to execute all of the introduced algorithms, and for compilation, g++ is used. All experiments are conducted on a GNU/Linux machine, using Intel Core i3. With a timeout of 500 s, we have run every instance. These graphs are generated using the OriginPro-2018 version. Numerous instances are generated for diverse trip range and line number, then compared with RA, EA, and FSCF algorithms to analyze the performance of each of these. Here, for any selection, RA chooses the drivers or vehicles randomly from the available list. EA starts choosing from the top of the list and does not repeat any drivers or buses until it chooses everyone at least once. FSCF also chooses from the top of the list, but it repeats drivers or vehicles if the first chosen ones complete their trips and are again present in the available list. We have generated a variety of instances of this problem by fluctuating the parameters. Here, we assume each line's trip round time roughly includes mild congestion time. However, trip planning for any delays or missed trips is not considered due to any mishaps and accidents. We assumed every trip completes on time without any hindrance. The performance metrics are discussed in Table 4 briefly.

Figures 2a, b sequentially show the driver and vehicle number comparison of three discrete algorithms with a different number of trips where the number of lines is fixed to five. In the graph, the x-axis represents the number of trips and the y-axis



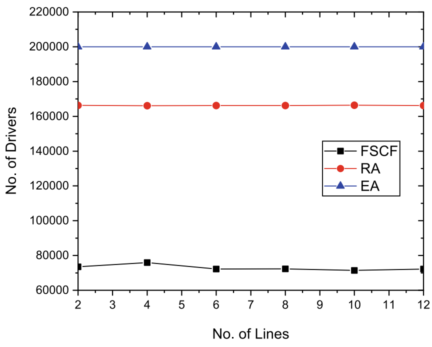
(a) Driver (b) Vehicle

(c) Runtime

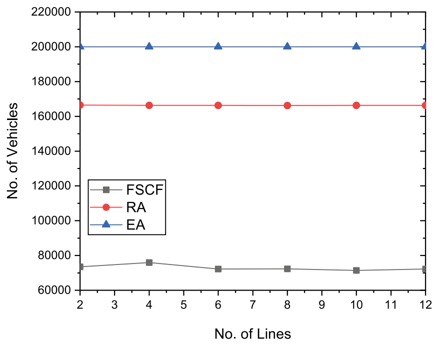
Fig. 2 Impact of number of trips (5 lines)

represents the number of drivers in Fig. 2a and the number of vehicles in Fig. 2b. It is quite noticeable that the FSCF algorithm gives better results from the beginning, and it improves when trips are increased. FSCF algorithm needs fewer number drivers and vehicles to complete trips where RA and EA need strikingly more. The runtimes of the three algorithms are depicted in Fig. 2c. Here, the *x*-axis is similar to the previously mentioned graphs, and the *y*-axis is runtime in milliseconds. Though FSCF gives better outcomes in drivers and vehicles selection, its execution time is usually a little higher than RA and EA.

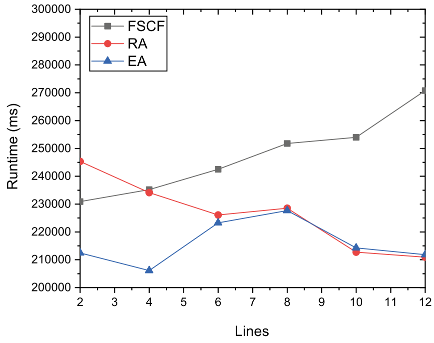
Now, we have fixed the trip numbers to 200,000 and then compared the driver number and vehicle number orderly in Fig. 3a, b and analyzed the performance of each algorithm. In both graphs, the *x*-axis represents the number of lines and the *y*-axis represents the number of drivers in Fig. 3a and the number of vehicles in Fig. 3b. Usage of drivers and vehicles does not seem to increase when line numbers are increasing; moreover, it maintains a nearly straight line for every algorithm. So in both cases, escalation of line number brings fewer changes in the driver and vehicle



(a) Driver



(b) Vehicle



(c) Run-time

Fig. 3 Impact of number of lines (200,000 trips)

selection. As mentioned in previous graphs, the FSCF algorithm gives better results in these cases when selecting drivers and vehicles.

There is also another runtime comparison in Fig. 3c of three algorithms with a different number of lines where trips are fixed to 200,000. In the graph, the x-axis represents the number of lines and the y-axis represents the runtime in milliseconds. Though in Fig. 3c runtime of all the algorithms has irregularities in the curvature.

5 Conclusion

The purpose of our work is to create a consolidated framework for bus-driver allocation according to a schedule that is designed by considering the usual rush-hour time in Bangladesh. In this paper, a well-organized timetable for bus transit is outlined to generate required trips per line at every time gap according to traffic peak hours. After generating the schedule dataset, it is used in three different methods, namely First Serve Check First (FSCF), Equal Allocation (EA), Random Allocation (RA) for scheduling the trips with a minimum number of buses and drivers. These algorithms are executed for a varied span of total trips and line count. The experimental results indicate that the FSCF method can schedule with a minimum number of vehicles and drivers, though the algorithm runtime is slightly higher than other approaches.

In this work, we did not consider any missed or delayed trips in the assignment of bus or drivers. In the future, we plan to develop a quality-aware driver selection mechanism while considering the uncertainties in the trips. Moreover, state-of-the-art heuristic methods can be compared to determine the effectiveness of the proposed methods.

Acknowledgements This work is supported by a grant for the "Research Fellowship (2020–2021)" funded by the Information and Communication Technology Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh.

References

1. Population of the dhaka. https://en.wikipedia.org/wiki/Dhaka. Accessed 2021-04-21
2. Rahman, A., et al.: Business model for bus operation in Dhaka city under proposed brt system scenario (2017)
3. Mahmud, K., Gope, K., Chowdhury, S.M.R.: Possible causes and solutions of traffic jam and their impact on the economy of Dhaka city. J. Mgmt. Sustain. **2**, 112 (2012)
4. Huisman, D., Freling, R., Wagelmans, A.P.: Multiple-depot integrated vehicle and crew scheduling. Transp. Sci. **39**(4), 491–502 (2005)
5. Li, H., Wang, Y., Li, S., Li, S.: A column generation based hyper-heuristic to the bus driver scheduling problem. Discrete Dyn. Nat. Soc. **2015** (2015)
6. Ma, J., Ceder, A., Yang, Y., Liu, T., Guan, W.: A case study of Beijing bus crew scheduling: a variable neighborhood-based approach. J. Adv. Transp. **50**(4), 434–445 (2016)
7. De Leone, R., Festa, P., Marchitto, E.: A bus driver scheduling problem: a new mathematical model and a grasp approximate solution. J. Heurist. **17**(4), 441–466 (2011)
8. Li, J., Kwan, R.S.: A fuzzy genetic algorithm for driver scheduling. Euro. J. Oper. Res. **147**(2), 334–344 (2003)
9. Chen, S., Shen, Y.: An improved column generation algorithm for crew scheduling problems. J. Inf. Comput. Sci. **10**(1), 175–183 (2013)
10. Borndörfer, R., Schlechte, T., Weider, S.: Railway track allocation by rapid branching. In: OASIcs-OpenAccess Series in Informatics, vol. 14. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2010)
11. Constantino, A.A., de Mendonça Neto, C.F.X., de Araujo, S.A., Landa-Silva, D., Calvi, R., dos Santos, A.F.: Solving a large real-world bus driver scheduling problem with a multi-assignment based heuristic algorithm. J. Universal Comput. Sci. **23**(5), 479–504 (2017)
12. Ibarra-Rojas, O.J., Delgado, F., Giesen, R., Muñoz, J.C.: Planning, operation, and control of bus transport systems: a literature review. Transp. Res. Part B Methodol. **77**, 38–75 (2015)
13. Gaffi, A., Nonato, M.: An integrated approach to ex-urban crew and vehicle scheduling. In: Computer-Aided Transit Scheduling, pp. 103–128. Springer (1999)
14. Dhaka transport coordination. http://dtca.portal.gov.bd/site/page/ace76617-c61e-4e60-86ee-8d64b81f6fba/-. Accessed 2021-09-21
15. Boyer, V., Ibarra-Rojas, O.J., Ríos-Solís, Y.Á.: Vehicle and crew scheduling for flexible bus transportation systems. Transp. Res. Part B Methodol. **112**, 216–229 (2018)

Emerging Applications

Incongruity Detection Between Bangla News Headline and Body Content Through Graph Neural Network

Md. Aminul Haque Palash, Akib Khan, Kawsarul Islam, MD. Abdullah Al Nasim, and Ryan Mohammad Bin Shahjahan

Abstract Incongruity between news headlines and the body content is a common method of deception used to attract readers. Profitable headlines pique readers' interest and encourage them to visit a specific website. This is usually done by adding an element of dishonesty, using enticements that do not precisely reflect the content being delivered. As a result, automatic detection of incongruent news between headline and body content using language analysis has gained the research community's attention. However, various solutions are primarily being developed for English to address this problem, leaving low-resource languages out of the picture. Bangla is ranked 7th among the top 100 most widely spoken languages, which motivates us to pay special attention to the Bangla language. Furthermore, Bangla has a more complex syntactic structure and fewer natural language processing resources, so it becomes challenging to perform NLP tasks like incongruity detection and stance detection. To tackle this problem, for the Bangla language, we offer a graph-based hierarchical dual encoder (BGHDE) model that learns the content similarity and contradiction between Bangla news headlines and content paragraphs effectively. The experimental results show that the proposed Bangla graph-based neural network model achieves above 90% accuracy on various Bangla news datasets.

Keywords Graph neural network · Low-resource language · Bangla news headline incongruity

Md. A. Haque Palash · K. Islam
Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh

A. Khan (✉)
Brac University, Dhaka, Bangladesh
e-mail: akibkhan0147@gmail.com

MD. A. Al Nasim · R. M. Bin Shahjahan
Pioneer Alpha, Dhaka, Bangladesh
e-mail: info@pioneeralpha.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_27

1 Introduction

News that is misleading or deceptive has become a major social issue. Much of the information published online is unverifiable, exposing our civilization to unknown dangers. Every day, the amount of news content produced skyrockets. However, unlike newspapers, which only print a certain amount of content each day, making articles online is relatively inexpensive. Additionally, many of these news stories are generated by automated algorithms [5], lowering the cost of news production even more. Several news organizations aim to attract readers' focus by employing news headlines unrelated to the main content to draw traffic to news stories among the competitions. News headlines are well-known for forming first impressions on readers and, as a result, determining the contagious potential of news stories on social media. People in information-overloaded digital surroundings are less inclined to read or click on the entire content, preferring to read news headlines. As a result, deceptive headlines may contribute to inaccurate views of events and obstruct their distribution. The headline incongruity problem in Bangla news is addressed in this study, which involves determining if news headlines are unrelated to or distinct from the main body text. Figure 1 depicts a scenario in which readers could expect to learn precise information about picnic places and picnic spot traders based solely on the headline; however, the news content comprises a Bangla movie advertisement. Because the body text is only accessible after a click, many readers will ignore the discrepancy if they only read the news headlines. Inconsistency in content is becoming more of a concern, lowering the quality of news reading.

Researchers have suggested a number of realistic techniques to address the detection problem as a binary classification utilizing deep learning based on manual annotation (i.e., incongruent or not). A neural network technique is used to learn the features of news headlines and body text in a new way [6]. These techniques, however, face two significant obstacles. For starters, current models focus on recognizing the link between a short headline and a lengthy body text that can be thousands of words long, which makes neural network-based learning difficult.

Second, the lack of a large-scale dataset makes training a deep learning model for detecting headline inconsistencies, which involve a variety of factors, more difficult. The headline incongruity problem is solved using a Bangla graph-based hierarchical dual encoder (BGHDE) in this study. It captures the linguistic interaction between a Bangla news headline and any size of body text. By integrating the headline and body paragraph text content as nodes, it makes use of the hierarchical architecture of news stories. On one hand, this method builds a network with nodes for headlines, and on the other hand, it creates a graph with nodes for body paragraphs. Then, between these nodes, we link undirected edges.

The BGHDE is trained to calculate edge weights on a large number of headline and paragraph nodes, with the more relevant edge weight being given. Then, by aggregating information from surrounding nodes, BGHDE updates each node representation. The iterative update technique propagates relevant data from paragraph nodes to the headline node, which is important for spotting content inconsistencies.

In this paper, we basically followed the data preparation and modeling part from the paper [21] and tried to reproduce the work for Bangla. However, our contribution can be summarized as below:

1. We propose a graph-based architecture for the first time to detect incongruity between Bangla's new headline and body text.
2. We provide an approach for synthetic data generation and made all the code and preprocessed dataset publicly available.[1]
3. We have analyzed and presented our model performance both on synthetic dataset and real-world datasets. Apart from news dataset, we also tested our model performance, how it can detect irrelevant comments on manually collected comments from social networking cites like Facebook and YouTube.

The findings demonstrate that the proposed method can be utilized to detect inconsistencies in real-world Bangla news reports. As shown in Fig. 1, even for new themes like picnic locations and traders, BGHDE can successfully detect anomalies in headlines and body content. The remainder of this article will be discussed in the following manner: Sect. 2 begins with a summary of the research on headline incongruity detection and the use of neural graph networks with text. The data creation procedure is then introduced in Sect. 3. Finally, Sect. 4 discusses the baseline models that were tested in this work. The proposed model is then thoroughly discussed. In Sect. 5, we give the experimental setup for model evaluation, as well as a discussion of the results obtained from various kinds of Bangla news and empirical study in the field. Finally, we consider the study's implications in terms of combating online infodemics and news fatigue. Finally, Sects. 6 and 7 conclude the paper with a discussion of the study's shortcomings as well as future research opportunities in the field of news incongruence detection.

2 Related Works

In the present era of the digital world, people are more likely to skim through the headlines and perceive information using both direct memory measures and more indirect reasoning measures [3]. Thus, misleading information can lead to cause more harm. There have been a lot of machine learning algorithms used. Despite the lack of a large-scale realistic dataset being the main issue with this challenge, some studies were able to obtain good results by utilizing manually annotated small-scale datasets. An attention-based hierarchical dual encoder model is used to detect incongruity between the headline and the news body [22]. They have also published a million-scale dataset and introduced augmentation techniques to enhance the training data.

Both incongruity detection and stance detection are related as they share a common basis: to identify the relationship between a brief piece of content and a long piece of

[1] https://github.com/aminul-palash/bangla-news-incongruity-detection.

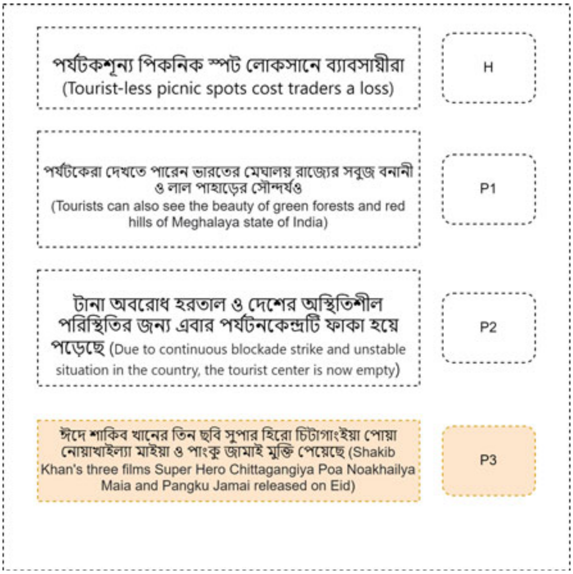


Fig. 1 Graph illustration of an incongruent headline problem between the headline and the body paragraphs. The inconsistency between paragraph 3 and other texts is represented by a separate color node in the graph

content. The goal of the 2019 Fake News Challenge was to encourage the development of stance detection models. For false news identification, a multi-layer perceptron model with one hidden layer [14] propagated lexical and similarity features such as bag-of-words (BOW), term frequency (TF), and term frequency-inverse document frequency (TF-IDF). However, the winner of the contest for Fake News Challenge 2019 used the XGBoost [2] algorithm with extracted hand-crafted features. One of the recent studies used semantic matching [10] dependent on inter-mutual attention by generating synthetic headlines that corresponded to the news body content and original news headline to detect the incongruities. However, for low-resource languages like Bengali, this kind of work is rare due to a lack of well-developed datasets. In the paper [6], they explored neural network models and pre-trained transformer models to detect fake news for fake news detection in Bangladesh. They have also released an annotated Bangla news dataset that can be used to create an automated fake news detector.

Graph neural network (GNN) is semi-supervised learning which utilizes graph-structured data [8]. GNN models can learn hidden-layer representations by encoding local graphs and features of vertices while maintaining the linear model scales of graph edges. The fundamental advantage of GNN models over traditional models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) is that GNN models embed relational information and pass it on to neighbor nodes during training. As a result, GNN models have succeeded miraculously in NLP tasks like question-answering [17, 20], relationship extraction [24], and knowledge base completion [16].

Fake news detection using GNN models has become a common practice nowadays. The paper [23] introduced FAKEDETECTOR, an automatic fake news detector model based on explicit and hidden text features that constructs a deep diffusive network model to simultaneously learn the representations of news articles, producers, and subjects. The hierarchical graph attention network was also used by the researchers, which uses a hierarchical attention method to identify node representation learning and then employs a classifier to detect bogus news [13].

To the best of our knowledge, our work is the first to detect incongruity between news headline and body for Bangla. In our work, we proposed a Bangla graph-based hierarchical dual encoder (BGHDE) model for automatic detection of Bangla news headline and body.

3 Datasets for Detecting Headline Incongruence Problem

We propose an approach for detecting incongruity between Bangla news headlines and content where we specifically tackle three significant challenges for preparing the dataset. We followed the same approach from this paper [21] for preparing our dataset for Bangla language. The first is the scarcity of manually annotated training datasets, as well as the high expense of creating them. The second is the length of news stories, which can often be long and arbitrary, making them challenging to model for machine learning. The last one is the paragraph creation from the Bangla news corpus, as our news dataset does not contain any paragraph separation. In the sections below, we will go over each obstacle in detail.

There are millions of news articles over the internet, and previously, the ground truth was manually annotated in earlier investigations [1, 19]. However, it is almost impossible to annotate ground truth for each news manually. Therefore, although some previous studies created manually annotated datasets, we adopt an automated approach for creating the annotations.

To begin with, we gather news stories from reputable target news sources. Then, we pick a collection of target news articles at random to manipulate, i.e., change them to look like incongruent news sources.

For the news pieces chosen for alteration, we replace some paragraphs with paragraphs obtained from other news sources. This collection of parts is created individually and is not used for training or testing. We carefully monitor the alteration process to ensure that no news stories are duplicated.

The news corpus that we employ in this paper comes from the famous Bangla news site Prothom Alo,[2] which consists of over 400k authentic Bengali news articles.

Figure. 2 shows the overall workflow of incongruent level data preparation [21]. There is one target news article and one sample news article. In our case, we select sample news articles randomly from the real news corpus. The sample news content paragraphs are then blended in with the intended news content. The maximum number

[2] https://www.kaggle.com/furcifer/bangla-newspaper-dataset.

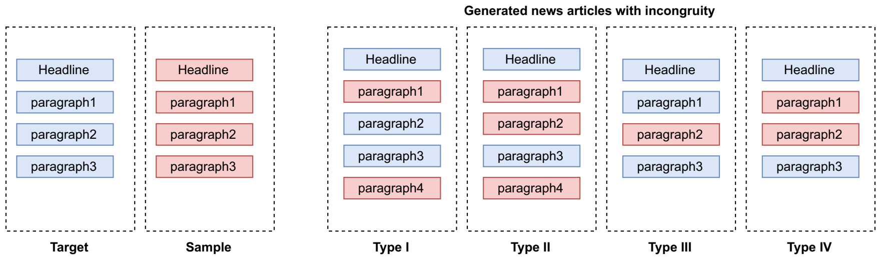


Fig. 2 Illustration of how news items with incongruent headlines are created. Paragraphs from samples articles are mixed with target news articles randomly. Total five types of data mix-up processes are shown (Type I, Type II, Type III, and Type IV)

Table 1 Statistics of our Bangla data

Dataset	Samples	Headline		Content	
		Avg.	Std.	Avg.	Std.
Train	228,000	5.58	1.45	319.35	205.41
Dev	120,000	5.58	1.43	319.01	241.06
Test	120,000	5.57	1.43	323.55	214.124

of paragraphs that can be switched is determined by the number of paragraphs in the target news article. We randomly manipulate paragraph swapping processes that address different difficulty levels on the target news article.

From Fig. 2, there is a total of four types of samples that can be generated by the generation process. But we do not pick the first two types (I and II) for preparing our training datasets so that we can keep consistent same distribution with the original and generated news content in length. So we keep types three and four (III and IV) for preparing our synthetic training datasets (Table 1).

As there was no paragraph separation on our collected dataset, we had to separate paragraphs synthetically based on the number of sentences on the news content. Thus, we split the news article by paragraph, where each paragraph contains five, ten, twenty, and so on sentences based on the length of news content.

Furthermore, we preprocessed the data by removing unnecessary punctuation from both headline and body content. We also discarded the news data that contains small body content.

4 Methodology

Our goal is to determine if the content of the news article matches the news headline. First of all, we extracted semantic information from the data as the distributed representation of words and sub-word tags has proven effective for text classification problems. Therefore, we used pre-trained Bangla word embedding, where an article is represented by the mean and the standard deviation of the word vector representation. For this experiment, we used Bengali GloVe 300-dimensional word vectors pre-trained embedding,[3] and our coverage rate was 43.34%.

For detecting news incongruities, we adopt a graph neural network (GNN)-based architecture following [21]. GNN is a type of deep learning method that can be used for processing data represented as a graph [15] for making node-level, edge-level, and graph-level analyses.

Our proposed Bangla graph-based hierarchical dual encoder (BGHDE) illustrated in Fig. 3 takes the headline and paragraph contents into account to detect incongruency in an end-to-end manner. Our proposed architecture is followed from this paper [21] which implements an incongruity detection method for English dataset. The architecture consists broadly of four steps which we describe briefly below.

First of all, the BGHDE creates a graph that is undirected representing its innate structure for each news article. After that, it was utilized to train the neural network with graph structure. Finally, a hierarchical gated recurrent unit (GRU)-based bidirectional recurrent neural network (RNN) framework is utilized to generate a representation with node structure of each headline text and paragraph text, as well as context-aware paragraph representation.

A group of nodes called vertex represents the headlines text and each corresponding paragraph of the Bangla news material. The edge of the graph represents the link between headlines and its corresponding paragraphs of the Bangla news.

Secondly, to avoid undesirable flattening of the node representation between the congruent and incongruent paragraphs during GNN propagation, the next step is to learn the edge weights of the input graph representation.

We employ the paragraph congruity value as a label to supervise the edge weights during the cross-entropy loss.

Because of this edge-level monitoring, the BGHDE can assign larger weights to congruent paragraphs and lower weights to incongruent paragraphs, allowing congruent paragraphs to communicate more information to the headline node than incongruent paragraphs alone.

Later, the node characteristics are transmitted into surrounding nodes in the third phase using the established graph structure and trainable edge weights from the GNN framework. In an edge-weighted form, BGHDE uses the convolutional graph network (GCN) aggregation function [9].

Finally, the incongruity of scores of news pieces is predicted for the last stage. The GNN graph classification problem is the same as this. The global-level graph representation and the local-level node representation must be fused together; BGHDE adapts a fusion block presented in [18].

[3] https://github.com/sagorbrur/GloVe-Bengali.

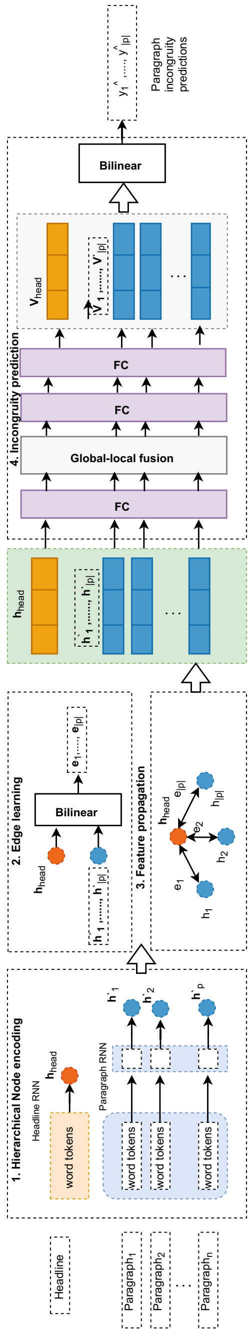


Fig. 3 Overview of our proposed model for Bangla news incongruity detection

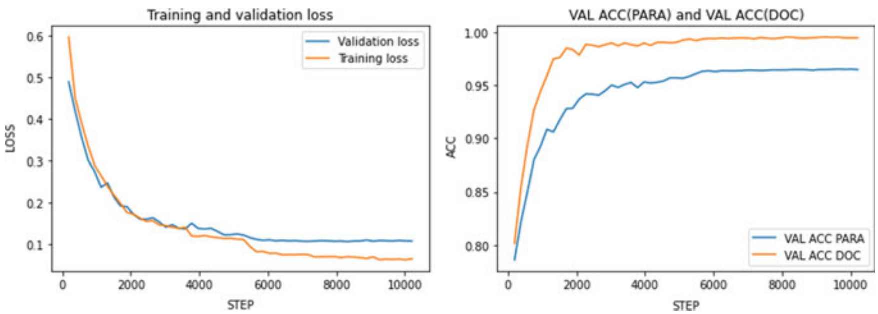


Fig. 4 Left plot represents training and validation loss, and the right plot represents accuracy of the validation data during training

5 Experiments

The title and the appropriate paragraphs of the subsequent body text are encoded using a single-layer GRU with 200 hidden units. In contrast, a single-layer bidirectional GRU with 100 hidden units is used to encode the paragraph-level RNN. There are three GNN layers, each with 200 hidden units. Hidden unit dimensions of 200, 200, and 100 for the FC layers are applied after feature propagation on the graphs, respectively. We use the Adam optimization algorithm [7] to train the model, starting with an initial learning rate of 0.001, which is decayed every three epochs by a factor of 10. We use 120 samples for each mini-batch during training. We clip the gradients with a threshold of 1.0. The hyperparameter for tradeoffs for edge loss is set to 0.1. We use pre-trained Bangla[4] GloVe [12] embedding consists of 300-dimensional vectors to initialize the word embedding. The number of words that appear at least eight times in the training dataset determines the size of the embedding matrix's vocabulary. The model has 1,214,702 total trainable parameters. We obtained an accuracy of 0.9560 and an AUC score of 0.9860 on the validation set. We show the loss and paragraph and document accuracy obtained during training in Fig. 4.

We use PyTorch [11] and PyTorch Geometric [4] frameworks to implement the model and Google Colab to run the experiments.

6 Results and Discussion

To validate our process, we conduct rigorous quantitative and qualitative analyses. We conducted a large-scale experimental analysis to evaluate our proposed Bangla graph-based hierarchical dual encoder (BGHDE). We performed an evaluation based on criteria like accuracy on both paragraph-wise and whole document or news article

[4] https://github.com/sagorbrur/GloVe-Bengali.

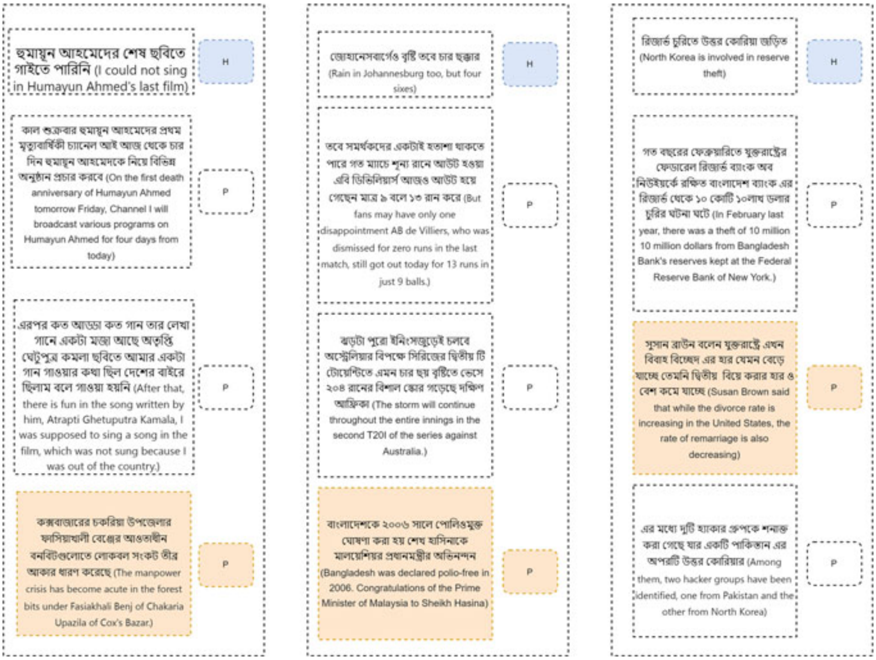


Fig. 5 Examples of some partial representation of news articles that contain incongruity between headline and content paragraphs, and colored paragraph is the incongruence paragraph

calculated while all types of a common evaluation matrix for the predictive model also added.

6.1 Performance Evaluation on Synthetic Dataset

Figure 5 illustrates some examples of incongruent data detected by our proposed model on synthetic test datasets. Our tested dataset contains news from both Bangladeshi[5] and West Bengal[6] news articles which helps to analyze the model performs better. Table 2 shows the model's performance on different datasets when it was used to detect headline inconsistencies.

[5] https://www.kaggle.com/ebiswas/bangla-largest-newspaper-dataset.

[6] https://www.kaggle.com/csoham/classification-bengali-news-articles-indicnlp.

Table 2 Second table

Dataset	Size	Acc (para)	Acc (doc)	**Evaluation**			
				Precision	Recall	$F1$-score	Support
Prothom Alo	6000	0.9658	0.9918	0.98.80	0.9956	0.9918	[3000 3000]
bdnews24	2000	0.9175	0.94.50	0.7554	0.913	0.9431	[1000 1000]
Ananda Bazar	1000	0.9175	0.9450	0.9623	0.97	0.9431	[500 500]
Ebela	5000	0.9192	0.9702	0.970	0.9704	0.9702	[2500 2500]
Zee News	5000	0.9026	0.9542	0.9511	0.9576	0.9543	[2500 2500]
Ittefaq	8000	0.9445	0.9866	0.9812	0.9922	0.9866	[4000 4000]
Jugantor	6999	0.9494	0.9862	0.9830	0.9893	0.9861	[3458 3477]

6.2 Evaluation on Real-World Dataset

To see how effective our dataset and proposed models are in detecting incongruent headlines in the real world, we have conducted this process by collecting data containing actual articles in which any form of the generation process has not modified the body text. It is difficult to annotate read news containing incongruity manually, and we perform inference on a real news dataset without annotations. After that, we evaluated our model performance manually. But we achieved very poor performance. For example, the datasets we used have no paragraph separation, and we need to separate the articles into paragraphs randomly. Another one is that we only train our model on synthetic datasets.

We also evaluated our proposed model on detecting incongruent comments Fig. 6 on different sites like YouTube, Facebook, and various Bangla news sites. As a

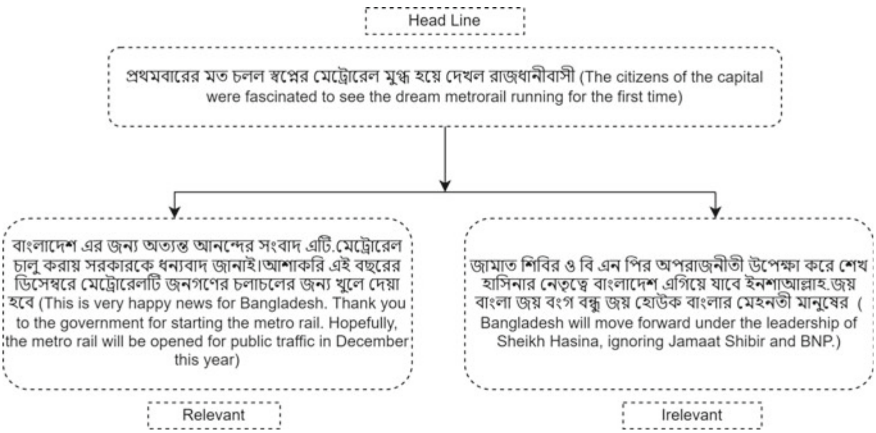


Fig. 6 Incongruity detection on comments data

result, we collected more than two hundred comments data containing relevant and irrelevant comments with corresponding news articles. We achieved an accuracy of 0.73 on the Bangla comments dataset.

7 Conclusion and Future Work

For the first time, a graph neural network was used to handle the headline incongruity problem in Bangla news stories. We discovered a few false positive scenarios when a model misinterpreted a coherent article for an incongruent headline using manual annotations. Although the computer accurately predicted the label, according to the idea of headline incongruity, such a "briefing" item does not mislead readers by delivering false information. Our findings suggest that more research is needed in the future to improve data generation and gathering processes. The findings of the evaluation experiments, however, reveal that the proposed technique accurately detects such deception. We hope that our research helps to create more trustworthy online news ecosystems in Bangla.

References

1. Chakraborty, A., Paranjape, B., Kakarla, S., Ganguly, N.: Stop clickbait: detecting and preventing clickbaits in online news media, pp. 9–16 (2016). https://doi.org/10.1109/ASONAM.2016.7752207
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp. 785–794. ACM (2016). https://doi.org/10.1145/2939672.2939785
3. Ecker, U.K.H., Lewandowsky, S., Chang, E.P., Pillai, R.: The effects of subtle misinformation in news headlines. J. Exper. Psychol. Appl. **20**(4), 323–35 (2014)
4. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. arxiv:1903.02428 (2019)
5. Guzman, A.L., Lewis, S.C.: Artificial intelligence and communication: a human-machine communication research agenda. New Media Society **22**(1), 70–86 (2020). https://doi.org/10.1177/1461444819858691
6. Hossain, M.Z., Rahman, M.A., Islam, M.S., Kar, S.: Banfakenews: A dataset for detecting fake news in bangla. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11–16, 2020, pp. 2862–2871. European Language Resources Association (2020). https://aclanthology.org/2020.lrec-1.349/
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. CoRR (2016). arxiv:1609.02907
9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2017)

10. Mishra, R., Yadav, P., Calizzano, R., Leippold, M.: Musem: Detecting incongruent news headlines using mutual attentive semantic matching. In: Wani, M.A., Luo, F., Li, X.A., Dou, D., Bonchi, F. (eds.) 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, Miami, FL, USA, December 14–17, 2020, pp. 709–716. IEEE (2020). https://doi.org/10.1109/ICMLA51294.2020.00116
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: an imperative style, high-performance deep learning library. arxiv:1912.01703 (2019)
12. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
13. Ren, Y., Zhang, J.: Fake news detection on news-oriented heterogeneous information networks through hierarchical graph attention. In: International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18–22, 2021, pp. 1–8. IEEE (2021). https://doi.org/10.1109/IJCNN52387.2021.9534362
14. Riedel, B., Augenstein, I., Spithourakis, G.P., Riedel, S.: A simple but tough-to-beat baseline for the fake news challenge stance detection task. CoRR (2017). arxiv:1707.03264
15. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Netw. **20**(1), 61–80 (2008)
16. Schlichtkrull, M.S., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.) The Semantic Web—15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings. Lecture Notes in Computer Science, vol. 10843, pp. 593–607. Springer (2018). https://doi.org/10.1007/978-3-319-93417-4_38
17. Song, L., Wang, Z., Yu, M., Zhang, Y., Florian, R., Gildea, D.: Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. CoRR (2018). arxiv:1809.02040
18. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds (2019)
19. Wei, W., Wan, X.: Learning to identify ambiguous and misleading news headlines, pp. 4172–4178 (2017). https://doi.org/10.24963/ijcai.2017/583
20. Xiao, Y., Qu, Y., Qiu, L., Zhou, H., Li, L., Zhang, W., Yu, Y.: Dynamically fused graph network for multi-hop reasoning. CoRR (2019). arxiv:1905.06933
21. Yoon, S., Park, K., Lee, M., Kim, T., Cha, M., Jung, K.: Learning to detect incongruence in news headline and body text via a graph neural network. IEEE Access **9**, 36195–36206 (2021). https://doi.org/10.1109/ACCESS.2021.3062029
22. Yoon, S., Park, K., Shin, J., Lim, H., Won, S., Cha, M., Jung, K.: Detecting incongruity between news headline and body text via a deep hierarchical encoder. CoRR (2018). arxiv:1811.07066
23. Zhang, J., Dong, B., Yu, P.S.: Fakedetector: Effective fake news detection with deep diffusive neural network. In: 36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20–24, 2020, pp. 1826–1829. IEEE (2020). https://doi.org/10.1109/ICDE48307.2020.00180
24. Zhang, Y., Qi, P., Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31–November 4, 2018, pp. 2205–2215. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1244

Computer Vision-Based Waste Detection and Classification for Garbage Management and Recycling

S. M. Yeaminul Islam and Md. Golam Rabiul Alam

Abstract Waste management systems and their inherent problems are still a matter of great concern amid this cutting edge of science and technology. Pile of untreated waste in open environment aids to a wide range of problems such as air and water pollution, untidy and unhealthy surroundings, wastage of recyclable materials, potential health risk of waste management workers and many more. The root cause of this problem points to a single fact that is too much manual labour involved in the garbage treatment process (collection, separation and recycling)—cannot keep up to the pace with which garbage generation happens. An efficient recycling method is imperative to solve this problem, which can be achieved by a fast, real-time garbage detection and classification system. In this research, we will propose a novel deep learning-based approach for automatic detection and classification of five kinds of waste materials, namely, *kitchen waste, glass waste, metal waste, paper waste, plastic waste*, from the garbage dump for an efficient recycling process, which not only improves the efficiency of the current manual approach but also provides a scalable solution to the problem. The contributions of this paper include a fully human labelled data set consists of 2200 images of garbage dump with 135,541 annotated objects from aforementioned categories and a real-time garbage object localization and classification framework based on a lightning fast, fairly accurate and end-to-end deep learning algorithm. For the baseline, we have articulated a single-stage object detection and classification framework and initialised the detector training process with pre-trained weights (trained on MS COCO data set) of the feature extractor. Then, with some fine-tuning and employing a few transfer learning tricks, we have proposed a waste object detection framework, that yields a mAP of 66.08% at an IoU threshold of 0.35 with an inference speed of roughly 55–58 ms in a single-GPU environment on both images and videos. It surpasses the performances of all the contemporary frameworks which deal with waste separation task.

S. M. Yeaminul Islam (✉) · Md. G. R. Alam
BRAC University, Dhaka, Bangladesh
e-mail: s.m.yeaminul.islam@g.bracu.ac.bd

Md. G. R. Alam
e-mail: rabiul.alam@bracu.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_28

Keywords Waste classification · Solid waste separation · Garbage recycling · Garbage waste data set · Object detection · Computer vision · Deep learning · Urban environment

1 Introduction

Bangladesh is a densely populated country with its population is on course to be around 17 crore by the end of this year. This huge population comes with a lot of human-made hazards, and 'solid waste' is surely a major part of it. An enormous volume of garbage waste is getting generated each day, particularly in urban areas, and regrettably, their treatment process is getting deteriorated as the days pass by. The major cities of Bangladesh generate around 16 thousand tons of waste a day, which summing up to more than 5.84 million tons a year. This amount is on the rise and will reach 47 thousand tons per day and approximately 17.16 million tons per year due to rapid expansion of urban area, population growth and hike in per capita waste generation rate. With the current amount of urban inhabitants, the waste generation rate is 0.41 kg per capita per day and the garbage collection efficiency ranges from 37 to 77% in different cities [1]. So, a big difference has been observed between the wastes being generated and the wastes being treated here. Waste materials are not properly handled or recycled as a result of a lack of awareness and implementation of the legislation. Consequently, it is a growing headache for the municipalities—how to manage these solid waste efficiently and ensure not only a healthy environment for the residents but also make effective use of the litters as well through recycling (Fig. 1).

Fig. 1 Challenges of vision-based garbage recycling system for recycling

1.1 Research Problem

The complete garbage management process through proper recycling and disposing takes huge amount time and manual labour. Recycling of solid waste is mostly dependant upon the garbage separation process, as different types of waste require different treatment. Automated solution to this problem involves computer vision and sensor-based approaches, which can help an automatic machine to filter out the waste objects as per their pre-defined classes and tag them accordingly. It imposes a wider range of challenges as well. The waste objects in nature can be obscured, twisted, camouflaged, transparent, aged and transformed in different forms, which makes it extremely difficult for such systems, to generalise well in production environment. In spite of several attempts like, letsdoitworld (https://github.com/letsdoitworld/wade-ai), the problems still persist in absence of an enriched data set with contextual images and a solid framework to detect and classify major categories of waste. So, our research problem is to find a fast and efficient waste object detection and classification system which has to work well in real-world multiple overlapping garbage objects and multi-class scenario that impose aforementioned challenges.

1.2 Aims and Objectives

A few decades back, a robotic agent, moving inside a dumping station and deliberately segregating wastes into pre-defined categories, will surely qualify for a science fiction movie! Fortunately, with the advent of IoT-enabled smart devices and robotics, fast and remote maintenance of mechanical activities can be facilitated and waste management can surely be a part of it. But, a problem arises that those systems require accurate representation of the kinds of waste to be treated. Some aspiring researchers addressed this problem and their solutions [2–6] but lack of contextual representation of garbage objects in their data sets, high inference cost, etc., make them questionable for edge device deployment. Now, detecting trash in open environment is a daunting task, as it can be twisted, pellucid, shattered and camouflaged and adding to that, the vast feature diversity of our natural world has to be taken into consideration as well. In this work, we will propose a solution of an automated garbage separation system using computer vision and end-to-end deep learning techniques, to encounter these issues. Our proposed system will be able to detect, localise and classify garbage objects from piles of waste in natural environment all in real time through pictures/videos/cameras. This information then can be passed to a robotic agent immediately to automate the process using a microcontroller-based embedded system. So, in summary, our propositions are

1. A fully human labelled data set consists of 2200 images of garbage dump. The data set should be evenly balanced to avoid data driven bias, which represent contextual images to fight against concept drift and data drift problem, annotated up to object-level granularity to detect and classify individual object from images.

2. A real-time garbage separation framework based on a fast and accurate object recognition algorithm, which can detect and classify multiple overlapping waste objects across three different scales (small, medium, large) from a variety of backgrounds.

1.3 Organisation of the Report

The report is structured in the following manner—In Sect. 2, we shall try to explain the concepts related to the subject matter in detail. In Sect. 3, a comprehensive walk through of the previous studies will be provided. Then, in Sect. 4, we shall briefly explain the data collection procedures and considerations of our proposed garbage detection data set. Afterwards, in Sect. 5, the framework architecture and design choices will be reviewed. Finally, we will wrap up with the experimental results and some future scopes of this research in Sects. 6 and 7, respectively.

2 Background Study

2.1 Waste Generation

Waste generation is something that cannot be avoided as long as we, humans, live in this beautiful planet Earth. Garbage has been the by-product of this ecosystem since the outset of human civilization. But during the recent years, there is a significant amount of artificial materials prevail around us, which, when converted to waste, result in dangerous non-biodegradable objects—which is the most concerning thing about waste generation nowadays. As types and quantities of waste vary in different localities, we will try to depict the scenario from a Bangladesh perspective. In city areas, MSW generation rate is increasing with the growth of population. As reported in [7], a total of 7.7 thousand tons of solid waste are being generated from the six divisional cities of Bangladesh on a daily basis, while the capital itself has a 69% share in the total waste volume. The formation of the total stream of waste is approximately 74.4% of organic matter, 9.1% of paper, 3.5% of plastic, 1.9% of textile and wood, 0.8% of leather and rubber, 1.5% of metal, 0.8% of glass and 8% of other wastes [8]. The constituents behind this waste composition are population growth, advanced life styles, economic conditions, seasonality, weather, amount of recycling and waste management programmes. The waste generation rate is linearly evolving with respect to the GDP growth of this nation as well.

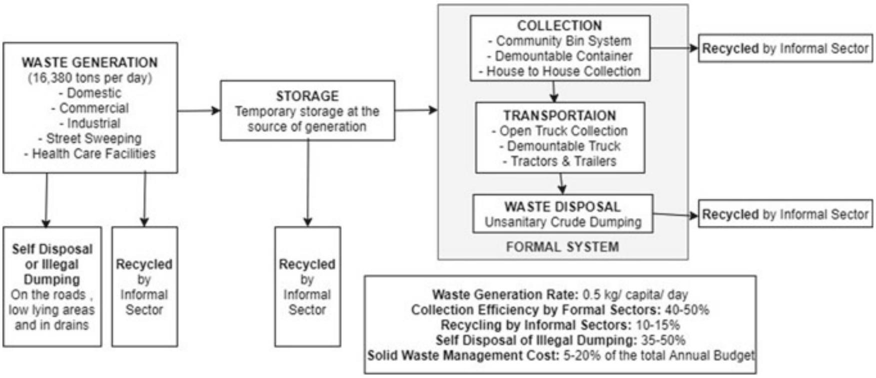


Fig. 2 Waste management framework in Bangladesh adopted from [1]

2.2 *Waste Management Systems*

The overall waste handling process here is not structured in a profound manner. Figure 2 is a clear illustration of the existing garbage handling framework in Bangladesh. There are three roughly established process of waste management in Bangladesh and the recycling process is completely ignored in almost all the cases (only 10–15% of total waste, that too by informal private channels), as the prerequisite waste separation task is a tedious thing to do.

2.3 *Sensor-Based Robotic Agent*

An autonomous robotic agent can have a vision sensor, to capture real-time scenes while surfing around the designated places. The captured images will then be analysed and waste object can be detected and classified via a waste object recognition framework. Then, the robotic agent will auto-separate the garbage objects in their respective categories for recycling or permanent disposal.

2.4 *Computer Vision*

Now, people, who use cameras as the sensors of the micro controllers, usually rely on images to give them information about the objects that they were looking at. In general, they use computer vision to identify objects within an image. The name computer vision is kind of self-explanatory which enables a computer to identify objects autonomously.

2.5 Object Detection and Classification

Image classification—Image classification algorithms take an image as input, transform it in a giant one-dimensional vector, process the input and yield the predicted label of that image with some metric like loss, probability, accuracy, etc. **Object detection—**Object detection algorithms take an image as input, transform it in a giant one-dimensional vector, process it and output one or more bounding boxes with the class labels, depicted on top of each bounding box.

Now, the object detection algorithms can be further classified into 2 major categories, namely **Single-Stage Detectors** refers to the family of algorithms that detects and classifies the object using a single network and one forward pass. This has attracted many researchers off late, mainly because of its faster inference speed and practical applicability. These models can also be optimised in such a way that they can be used in mobile devices. Examples—YOLO models [9–12], RetinaNet [13], SSD [14], etc. **Multi-Stage Detectors**, which employ two networks—one for predicting bounding boxes around detected objects and another for classifying the object in the box. These algorithms are based on region proposals and comparatively slow to make predictions than their single-stage counterparts. But they are more accurate than the single-stage detectors. Examples—R-CNN [15], Masked R-CNN [16], Fast R-CNN [17], etc.

2.6 YOLO Models

YOLO is a CNN-based object detection algorithm. This algorithm uses a single neural network and within a forward pass—the network breaks down the total image area into multiple local regions and predicts the bounding boxes and probabilities of those regions. The first version of this model [9] yielded greater localization errors, although operates at an impressive 45 FPS. In YOLOv2 [10], the authors adopted the idea of the anchor boxes from the R-CNN paper [17]. The representation of the bounding boxes is modified a bit (instead of estimating the bounding box position and size directly, offset values are predicted for moving and reshaping the pre-defined anchor boxes relative to a grid cell) to allow small changes which eventually expected to have a less dramatic effect on the predictions, resulting in a more stable model. The latest improvement YOLOv4 tried to combine all the common recipes of success from the rich vision recognition research domain termed as—*Bag of Specials*—Change in activation function, better backbone, better feature aggregator and propagation path, etc., and *Bag of Free bees*—lots of data augmentation techniques, modification in the loss function, using a better regularising method etc.; and present a fast, accurate and strong object detector which achieved state of the earth performance running on a single-GPU environment (Fig. 3).

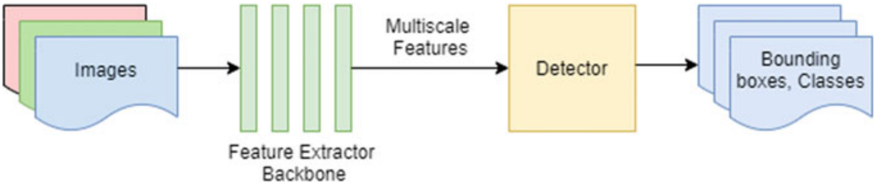


Fig. 3 High-level network architecture

3 Related Works

Our local researchers pointed out the problem and its severity in [1, 18]. In [18], authors proposed a high-level framework combining legal bindings, economic instruments and public awareness to tackle this problem. And in [1], authors portray the existing process of garbage collection to separation and processing scheme and expressed their proposition to haunt them down (mostly high-level proposition demanding process changes). In [8], authors proposed an electro-mechanical sensor-based system using microcontrollers and operational amplifiers to auto-separate common forms of wastes. But the practical applicability of this system is in question mainly because of the diversity of the waste objects possess which is difficult to capture with the amount of information that these sensors provide to make a prediction.

In [19], authors dealt with slightly easier problem which is tagging street cleanliness. They have used street cameras from different localities and tried to develop an array of localised models for area specific inference that indicates training a lot of discrete model which does not seem like a scalable solution of the problem. Moreover, they used Näıve Bayes and AdaBoost classifiers to make the prediction observing the whole image, which missed the detection granularity down to object level that is subjective for object separator to work. Recently, deep learning techniques have become increasingly popular because of their capabilities to extract rich features automatically from images, highly efficient parallel processing on GPUs to satisfy these resource hungry algorithms, fast and efficient optimizing techniques and a wonderful research community. That is why, most of the modern literature tried to address the problem using the same. In [20], authors used end-to-end deep learning-based system to do a similar task. For training efficiency, they have employed a pre-trained OverFeat model as the feature extractor backbone and GoogleNet model as classification head. Moreover, this paper addressed the detection granularity down to object level which makes it a better candidate for waste separator model. But, the data set they have developed is heavily biased towards, and did not handle overlapping objects, and didn't use techniques like non-max suppression to choose the most relevant bounding box from multiple predicted bounding boxes at a specific IoU threshold.

In [21], authors adopted a deep learning-based object localizer and classifier approach to figure out waste objects of 2 categories, namely biodegradable and

(a) (b) (c)

Fig. 4 Representative of waste images from [3, 22, 24]

non-biodegradable. But, authors did not portray any experimental results in this paper. No specific network architecture, loss function, result analysis or information regarding the data set are present here. In [22], authors adopted state of the earth MobileNet, DenseNet and inception networks for waste object classification and portrayed good level of accuracy metric. But, the waste object is not represented contextually (Fig. 4c) in the data set as they adopted the TrashNet data set and they have not taken the detection granularity to the object level as well. In [23], the authors used transfer learning to train a Faster R-CNN model pre-trained on PASCAL VOC data set which detects and classifies three categories of waste objects, namely landfill, recycling and paper. They have also adopted the TrashNet data set and engaged lots of augmentation techniques to artificially generate contextual images. This yields significant improvement over the last works involving this data set with a mAP of 68.3% but the algorithm does poorly in paper class and it also ignored some other waste categories like glass, metal, plastic, etc.

In [3], authors trained a multi-layer hybrid deep learning network. This system involves both high-resolution camera and sensors to extract useful features and finally processes the image features using a CNN-based algorithm and deploys a fully connected MLP to blend image features with other information to classify waste as recyclable or others. This system tried to mimic human sensory organs like eyes with a pre-trained AlexNet model and ears/noise with a MLP and achieved a better classification accuracy than other systems at that time. But 2 problems still persist—the model fails to pinpoint the waste object rather classifies the whole image frame which can contain multiple waste objects and although they employed image augmentation to present different form of the image to the model, and the background and natural diversity of the real world is still absent here as seen from Fig. 4a. In [24], the authors tried with classifiers like support vector machine and neural network models like AlexNet, with applied data augmentation techniques. But the context is missing here along with multiple overlapping object detection and classification which can be seen from Fig. 4b. An improved work using this data set has been published in [4], where authors tried data augmentation techniques like building collage of images, employing GANs, etc., to artificially manipulate the real-world scenario and yield some good results but no inference test has been provided involving real-world waste image, which raises the question of their practical applicability.

(a) Example 1 (b) Example 2 (c) Example 3

Fig. 5 Representative of waste images from [25]

In [25], authors used an end-to-end deep learning approach in a hierarchical manner to detect and classify waste in food trays. They have published their developed data set (Sample: Fig. 5) and adopted a multi-label multi-class classification approach to annotate the data set to improve the accuracy. But the usage of 7 individual CNNs to classify the waste objects significantly increases the computational complexity and inference time, making it hard to deploy the system in edge devices for practical applications. In [2, 5], authors used more realistic images to deal with context problems and both contributed their own data sets GINI and TACO but both of them represented very small amount of annotated objects (*1494* and *4784,* respectively) and unbalanced class distribution. Architecturally, [2] offers only binary detection—garbage is present or not which takes out the recycling application, and the inference time required is not feasible for practical applications as well. Reference [5] detects only litters leaving the biodegradable wastes out of equation. It does not support multi-scale detection; hence, small sized objects are often get ignored. Moreover, both of them used a region proposal-based 2-stage object detector which is slower to make inference on real-time data.

4 The Garbage Annotated in Context Data Set

Building the data set is one of the more challenging tasks of any machine learning project. The struggle becomes scarier when we do not have any benchmark data set for our problems—which actually happened in our case. Now, building data set can be done in a few different ways as follows. A brief overview of them is illustrated below.

4.1 Publicly Available Open Labelled Data Sets

The existing open source data sets regarding waste management domain use mostly out of context images, portray unbalanced class ratio, do not provide object-level annotations, etc. (as illustrated in previous section). Moreover, our required classes were absent in those data sets as well. That is why, we were unable to adopt any of them for our work.

4.2 Web Scraping

The next option is writing an automated web scrapper script to search keywords of required images and download accordingly. It is a proven method and has been adopted in various works like Web Scraping of COVID-19 news stories to create data sets (https://dl.acm.org/doi/fullHtml/10.1145/3453892.3461333). Now, images found on the Internet are of lots of different sizes and pixel resolutions. Some images are of pretty low resolution and our target task can contain 50–100 objects per frame which can make the job difficult for the detector on very low resolution images. So, we have written our own image-scrapper-from-web script (https://github.com/Yeaminul/Automation). After accumulating all the downloaded images, we also manually inspected each image and took decision as per the merit of the image to portray the real-world waste scenario. There is a demonstration of a few sub-categories of waste images that we have downloaded in Table 1. We have tried to cover the common categories of waste that are seen in the garbage dump in Bangladesh.

4.3 Capturing Images

Taking videos or capturing images of the subject can also be considered as an option to develop a data set of this kind. This can be achievable by (a) manually clicking each image or capture video by ourselves or (b) crowd sourcing, which is employing people to take images of the subject matter for us. (c) Installing programmed camera in our environment of choice. Due to lack of financial support and other endorsements, we were unable to take (b) and (c) as an option. So, we had to proceed with manually clicked photographs by ourselves.

Table 1 Search category granularity to accumulate garbage photos through web scrapping

Metal	Glass	Kitchen waste	Plastic	Paper
Alloy metal scrap	Broken glass jar	Biodegradable waste	Coke plastic bottle waste	Cardboard waste
Crushed coke cans	Broken glass waste	Chicken bones waste	Plastic waster bottle waste	Craft paper waste
Electronic scrap	Broken plate waste	Egg waste	Industrial plastic waste	Magazine paper waste
Ferrous metal scrap	Broken glass waste	Fish waste	Laboratory plastic waste	Newspaper waste
Household metal scrap	Crushed glass waste	Food waste	Medical plastic waste	Paper waste
Metal rod scrap	Glass bottle waste	Food waste close-up	Plastic bottle waste	Pizza box waste
Metal waste	Glass waste	Fruit waste	Plastic container waste	Paper scrap close-up
Non-ferrous scrap	Light bulb waste	Fruit waste close-up	Polythene waste	Shredded paper waste
Crushed steel drums	Recycled glass waste	Kitchen fish waste	7up plastic bottle waste	Soap packet waste
Tin canes waste	Scrap glass waste	Kitchen waste	Onetime coffee cup waste	Tissue waste
Used nuts and bolts	Shattered glass waste	Organic waste	Onetime plastic cup waste	Torn paper bag
	Broken window glass	Tea bag waste	Plastic bag waste	Used diapers
		Vegetable waste		Used tissue waste
				Waste paper junk

4.4 *Using Data Augmentation*

As deep learning algorithms are data hungry ones, when we have a relatively small data set, it becomes difficult to train a model and generalise it well on unseen examples. In these cases, different kinds of data augmentation can be adopted to extrapolate more training data. Geometric transformations like flipping, cropping, rotation and translation are the most common data augmentation techniques. Using data augmentation in training images not only enriches our data set by creating variation, but also reduces model over fitting problem. We have adopted a wide range of data augmentation techniques, which not only limited to the image transformations stated above, but also expanded towards some modern techniques like photometric distortion, cut mix, mix up and mosaic data augmentation as described in [12].

As we can see from Figs. 6 and 7, we have made our efforts clear to leave minimal distribution difference between the two sources of image collection, in easy words,

Fig. 6 Annotated images of garbage waste—self-taken photographs

Fig. 7 Annotated images of garbage waste—web scrapped images

we have tried to carefully pick up photographs form the scrapped images which can better mimic the real-world scenario.

4.5 Data Set Preparation Steps

Finally, if we summarise the steps that we had taken for building the data set,

1. First, we have visited 10 different dumping stations in Dhaka City; captured the spotted trash using mobile phones and accumulated more than 1200 images. We have taken images from different angles and verticals trying to mimic a moving agent which will automate the process. All the images are captured in different lighting conditions (Haze, Cloudy, Sunny, etc.) during daytime at an optimum resolution.
2. Then, we have carefully investigated the images and identified the varieties and different orientations with which waste appears in open environment.
3. After that, we have wrote a web scrapper in Python to download photographs of given category (refer to Table 1) from Google Images automatically. We have accumulated photographs of more than 70 different sub-categories waste, which belong to our main five categories.
4. From there, we have carefully investigated and picked 1000 photographs from the collection that matches the real scenarios by hand.

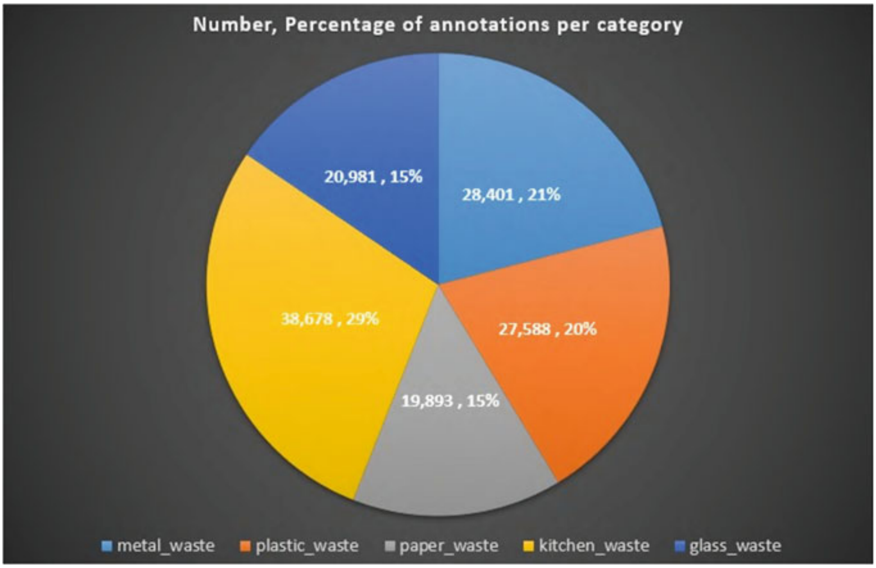


Fig. 8 Annotation count and percentage per category

5. We have annotated the final images using the labelling tool [26] in PASCAL VOC format.
6. The annotated images are then further verified by two another human individual and finally used for model development purpose. The primary objective had been to keep the contextual images only to guard against the data drift and concept drift problem.

We have also tried to make our data set evenly balanced so that no data driven bias exists in the trained model as illustrated in Fig. 8.

5 Model Overview

5.1 Problem Formulation

Considering an image of garbage waste, our goal is to predict and draw rectangular bounding boxes wrapped around each and every waste object present in that image and assign each detected object to one of the considered classes. Consequently, this can be formulated as a multi-label K-way classification problem. We employ an end-to-end deep learning method which uses a proven (1) feature extractor backbone to project rich image features to latent space, (2) leverage the idea of Spatial Pyramid Pooling and Path Aggregation Networks to significantly increase receptive field with almost no incremental time cost and collect image features from different level of the

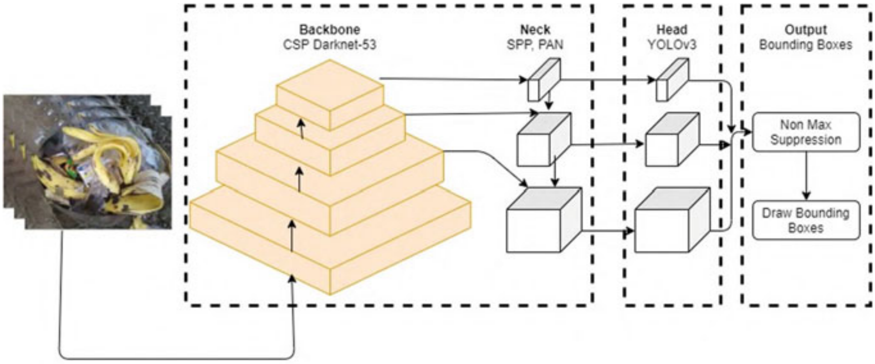


Fig. 9 CNN architecture used in our waste detection and classification framework

backbone network for different detector levels. We have also used a single objective function inclusive of all the losses and optimise that using Adam or RMSProp optimizer for dense prediction of waste objects. (3) Finally, we have used the YOLOv3 (anchor based) head, to detect and classify waste objects in three different scales. We have also used post-processing techniques like non-max suppression to filter out the most ideally fitted bounding box around the subject (Fig. 9).

5.2 *Feature Extractor Backbone*

While choosing the backbone of this framework, we had to make an optimum choice between maintaining both classification accuracy and detector efficiency simultaneously. The criterion of choice is—(1) higher input size—for multiple small object detection, (2) deep layers—for higher receptive field size to deal with the increased input size, (3) more parameters—for greater learning capacity of the model. The Darknet53 network used in [11] provided a significant improvement over the detection efficiency while maintaining good enough classification accuracy than the contemporary models. It shown promising results in cases of training the detector using the pre-trained backbone weights, which is very convenient for transfer learning strategy which we have adopted here. The CSPDarknet53 [27] that we have used here further optimises the framework by reducing the computations up to 20% hence decreasing the inference cost.

5.3 *SPP-PAN Neck*

Spatial Pyramid Pooling block after the CSPDarknet53 block significantly improves the receptive field size, highlighting the most significant contextual features without adding almost any computational overhead. Receptive filed size is an important property of the convolutional neural nets which enables a network to detect small to largest objects in multiple scale. We have also used *Path Aggregation Neck* as a parameter aggregator from different backbone level for different detection heads, to enable effective propagation of low-level image features towards the detector layers.

5.4 YOLOv3 Detection Head

This is the last block of the system and is responsible for detecting and classifying waste objects in images across 3 different scales having strides 32, 16 and 8, respectively. Consequently, an input image of size 416 * 416 will be able to make detections across scales 13 * 13, 26 * 26, and 52 * 52. Starting from input, up to the first detection layer, the network only downsamples the input, where feature maps of a layer with stride 32 are used to perform the detection. Afterwards, layers are upsampled by a factor of 2 and added with feature maps of a previous layer. Consequently, detections made at layer with stride 16. Similar strategy is repeated to yield the final detection at the layer with stride 8. Output processing techniques like *non-max suppression* also has been implemented in this block to filter the most accurate bounding boxes wrapped around an object from the predicted ones.

5.5 Loss Function

With the final detection result at hand, we will be able to calculate the loss with respect to the true labels now. The loss function consists of four parts: (1) centroid (*xy*) loss, (2) width and height (*wh*) loss, (3) objectness (object and no object) loss and (4) classification loss. Putting everything together in perspective, the final equation looks like below

$$\begin{aligned}\text{TotalLoss} = {} & \lambda * (\text{CIoU}((x, y), (x, y) * \text{obj mask}) \\ & + \lambda * (\text{CIoU}((w, h), (w, h) * \text{obj mask}) \\ & + (\text{BCE}(\text{class}, \text{class}) * \text{obj mask}) + (\text{BCE}(\text{obj}, \text{obj}) * \text{objmask}) \\ & + \lambda_{\text{noobj}} * (\text{BCE}(\text{obj}, \text{obj}) * (1 - \text{obj mask}) * \text{ignoremask})\end{aligned}$$

The first two components of the above function calculate the loss of the bounding box regression problem. Previously, RMSE had been used in this case but it does not converge well (when the large sized objects cause large loss values). IoU loss (refers to the ratio of overlapping area of the predicted BBox and union of them), being a ratio by nature, solves this problem. We have used a more recent advancement named as CIoU loss which brought the distance between centre points, and the aspect ratio into the loss calculation. Then, the usage of Binary Cross Entropy loss to calculate *objectness* and *noobjectness* loss is self-explanatory being a binary classification problem. By estimating this objectness loss, we can steadily show the model to distinguish an area of intrigue. At the same time, we do not need the algorithm to cheat by making wrong region proposals. Henceforth, we need noobjectness loss to punish those false positive assumptions. Finally, we used independent logistic classifiers instead of Softmax as suggested in [11] to calculate the class loss.

6 Implementation and Ablation Study

6.1 Implementation Details

Below are the key highlights of our training process.

- The weights of our feature extractor backbone have been initialized using the pre-trained model, CSPDarknet53 [27], which has been trained on huge MS COCO data set [28]. As a result, our model was able to exploit the rich hierarchy of already learned representations; helping it to reduce over fitting and achieve a better generalization. The original architecture has been tweaked to perform 5-way classification task as per our requirement. We have observed a significant uptake in performance with transfer learning through qualitative and quantitative experiments.
- We have adopted a fivefold cross-validation method for all the experiments and trained the detection head, freezing the weights of the feature extractor layer, to iteratively search the optimum set of hyperparameters which yield best training and validation results. For each fold, the data set is split randomly into 80% training set, 10% development set and 10% test set.
- As we have used an anchor-based method for the object detector, the dimensions of those priori boxes are vital. We have run a K-means algorithm on our whole data set images, to figure out the clusters of bounding box dimension distribution and found 9 optimal anchor box coordinates (3 per each scale). Hypothetically, we can say, it would increase the accuracy of the detector as we have considered IoU loss for the bounding box regression task (Fig. 10).
- We have used Tensorflow 2.0 for model building and training purpose. We have also exported the model using TFlite and TFServing for edge device deployment in real world use cases.
- The hyperparameters are as follows: the training steps are 10,000; the step decay learning rate scheduling strategy is adopted with initial learning rate 0.01 and multiply with a factor 0.1 at the 8000th steps and the 9000th steps, respectively; the network is optimized using RMSprop with momentum and weight decay are, respectively, set as 0.9 and 0.0005. Images of input resolution 416 * 416 have been fed to the network during the forward passes. All architectures use a single GPU to execute multi-scale training in the batch size of 64, while mini-batch size is 4 because of the huge architecture and GPU memory limitation (Studies

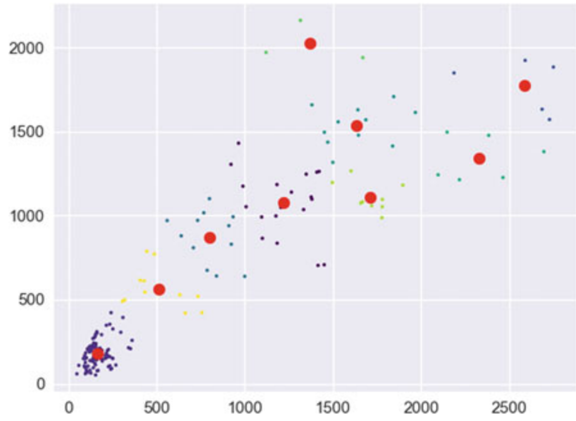


Fig. 10 Customised anchor box coordinates determined by running K-means algorithm over the whole data set images

from [12] revealed that mini-batch size has a pretty limited effect on detectors performance).

- Different experiments have been carried out borrowing ideas from [12] which involve using a lot of image augmentation techniques, cross mini-batch normalization, varying the IoU threshold for assigning ground truth, etc.
- All the training and experiments are run on a single-GPU environment using 4 GB GeForce GTX 1050 Ti GPU.

6.2 Overall Performance

Mean Average Precision (mAP) [29] has been used for model performance evaluation along with sensitivity (recall), F1 score and inference speed. The mAP score has been calculated by taking the mean AP over all the classes and compared at different IoU thresholds to extract the best setting. Finally, the IoU threshold has been set as 0.35 which achieved a good balance between sensitivity (recall) and specificity (precision). The defined metrics have been monitored via Tensorboard tool. Table 2 gives the performance of 2 Single-Stage Anchor box-based object recognition frameworks fine-tuned for our use case.

It is evident that the modified YOLOv4 model exceeds the earlier one is almost all the categories as it embarks upon all the bag of tricks from rich computer vision research domain. The 3 scaled approach makes the objects across various scales detectable. The deep architecture of the feature extractor has increased the size of the receptive field along with more parameters and robust learning. In both cases, the transfer learning strategy worked out well as the relatively small data set yielded considerable results. We have exported the model via TF Serving for edge device deployment and test sample inference. It requires 52 MB additional memory allocation for processing a single example. The relatively poor performance of glass waste and paper waste can be overcome by hard negative mining of the falsely detected

Table 2 Comparison of the speed and accuracy of 2 experimented frameworks on our waste data set

Method network	Inference speed (ms)	Sensitivity (%)	F1 score	AP (%)	AP (paper) (%)	AP (plastic) (%)	AP (glass) (%)	AP (kwaste) (%)	AP (metal) (%)
YOLOv3 Darknet53-FPN	70	40.47	0.47	57.3	47.30	47.02	63.40	60.06	68.70
YOLOv4 CSPDarknet53-SPP-PAN	58	62.25	0.65	66.08	57.55	65.67	57.32	73.06	76.80

examples. The overall performance of the model can be improved by enriching the data set to deal with more distribution differences of waste in wild.

6.3 Ablation Study

We have collected 50 more images outside the main data set from different places and conducted the ablation study. Firstly, we can see the effective low-level feature learning from the backbone in Fig. 11. This indicates the model generalises well on unseen examples even from slightly different data distribution which points towards effective **transfer of knowledge** from pre-trained weights. The use of **SPP-PAN Neck** instead of Feature Pyramid Neck results significant improvement over mAP score backed by better propagation of low-level features to high-level latent space and rise in receptive field resolution. The choice of **YOLOv3 head** as the Single-Stage Anchor-based detector makes the real-time inference possible by the making all the estimations in one forward pass. The choice of **CIoU** loss over RMSE significantly speeded up the convergence and reduced the training time. Our proposed **GACO data set** featuring multiple overlapping objects in contextual waste images annotated up to object-level granularity makes the model to operate well in real-life scenarios as well (Fig. 12).

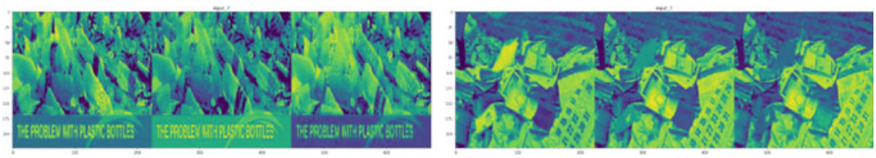


Fig. 11 Feature visualisation from feature extractor backbone

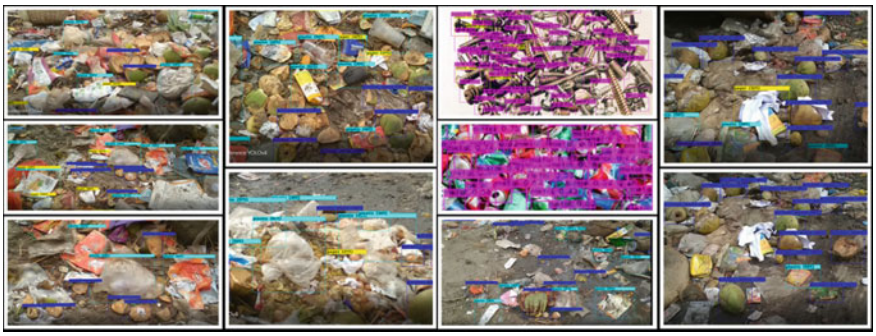


Fig. 12 Sample inference on real-world garbage data

Table 3 Bench-marking our 'garbage annotated in context (GACO)' data set w.r.t other data sets

Data set	#Images–#Objects	#Classes class	Distribution	Context
Garbage in images (GINI) [2]	1484–1484	2	Unbalanced	Real world
TrashNet [24]	3000–3000	6	Balanced	Not real world
Trash annotated in context (TACO) [5]	1500–4784	25	Unbalanced	Real world
Labelled waste in wild (LWW) [25]	1002–7200	19	Unbalanced	Real world
Garbage annotated in context (GACO)-ours	2200–135,541	5	Balanced	Real world

6.4 Result Comparison

In Table 4, we evaluate the proposed method and other prevailing waste object recognition methods with respect to mAP score, the type of problem addressed, sensitivity, inference time and deployment readiness in edge devices. All the data of other methods are directly obtained from their original papers. We have only considered the works which addressed waste recognition up to object-level granularity as hypothetically we can say that, garbage cannot be separated if not detected properly. Moreover, most of the models are trained and evaluated on imbalanced data sets containing manipulated waste images as described in Related Works, which arise the question of general applicability of those methods. So, we can say, our objective of delivering a fast, accurate object detection model capable of recognising common categories of wastes from messy real-world scenarios has been served through our proposed method. Table 3 gives the high-level comparison of our data set with respect to others.

7 Conclusion

So, here, we have proposed a deep learning-based single-stage waste recognition framework which can make real-time predictions of garbage objects in contextual images both in images and videos. We have taken a data-centric approach to build up a relevant data set and leverage the fine-grained ideas and recipes to build a decent framework for waste object detection and classification task. It is pretty lightweight, computationally cheap and easily deploy-able in edge devices which makes it convenient to use in practice. There is still a huge room for improvement in this work. The diverse distribution difference of the domain along with limited no. of training examples has been tried to cover up by transfer learning, data augmentation techniques up to some extent. Further research can be directed towards enriching the data set using

Table 4 Bench-marking the performance of our framework w.r.t the contemporary works

Model	Data set	#Annotation	Multi-label multi-class	Sensitivity	mAP	Inference speed	Readiness deployment
OverFeat-GoogleNet [20]	Own	4338	✓–✓	83%	63%		
Faster R-CNN-CNN [25]	LWW	7200	✓–✓		73.6%		
Masked R-CNN [5]	TACO	4784	✓–✓		26.10%		
Faster R-CNN [4]	TrashNet	8000	✓–✓	87%	81%		
CSPDarknet53-SPP-PAN-YOLOv3 (ours)	GACO	135,541	✓–✓	73%	66.08%	58 ms	✓

All the data are taken from the published papers provided by the authors the techniques that has been elaborated in The garbage annotated in context data set optimises the model further to enhance the learning capability with fewer parameters involved. However, we have proposed a novel deep learning-based approach to fight against one of the most prominent problems which prevails in our country at the moment. So, we believe its solid baseline to take the work forward and enrich further.

References

1. Bahauddin, K., Uddin, M.: Prospect of solid waste situation and an approach of environmental management measure (EMM) model for sustainable solid waste management: case study of Dhaka city. J. Environ. Sci. Nat. Res. **5**(1), 99–111 (2012)
2. Mittal, G., Yagnik, K.B., Garg, M., Krishnan, N.C.: Spotgarbage: smartphone app to detect garbage using deep learning. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 940–945 (2016)
3. Chu, Y., Huang, C., Xie, X., Tan, B., Kamal, S., Xiong, X.: Multilayer hybrid deep-learning method for waste classification and recycling. Comput. Intell. Neurosci. **2018** (2018)
4. Kulkarni, H.N., Raman, N.K.S.: Waste object detection and classification (2019)
5. Proen¸ca, P.F., Sim˜oes, P.: Taco: trash annotations in context for litter detection. arXiv preprint arXiv:2003.06975 (2020)
6. Vinodha, D., Sangeetha, J., Sherin, B.C., Renukadevi, M.: Smart garbage system with garbage separation using object detection
7. Concern, W.: Bangladesh waste database 2014, vol. 29, p. 2019. Dhaka, Waste Concern. Retrieved June 2014
8. Russel, M.M., Chowdhury, M., Shekh, M., Uddin, N., Newaz, A., Mehdi, M., Talukder, M.M.M.: Development of automatic smart waste sorter machine (2013)
9. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, pp. 779–788 (2016)
10. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
11. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
12. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
13. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Doll´ar, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
14. Liu, W., Anguelov D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: European Conference On Computer Vision, pp. 21–37. Springer (2016)
15. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
16. He, K., Gkioxari, G., Doll´ar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
17. Girshick, R.: Fast r-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
18. Abedin, M.A., Jahiruddin, M.: Waste generation and management in Bangladesh: an overview. Asian J. Med. Biol. Res. **1**(1), 114–120 (2015)
19. Alfarrarjeh, A., Kim, S.H., Agrawal, S., Ashok, M., Kim, S.Y., Shahabi, C.: Image classification to determine the level of street cleanliness: a case study. In: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), pp. 1–5. IEEE (2018)

20. Rad, M.S., von Kaenel, A., Droux, A., Tieche, F., Ouerhani, N., Ekenel, H.K., Thiran, J.-P.: A computer vision system to localize and classify wastes on the streets. In: International Conference on Computer Vision Systems, pp. 195–204. Springer (2017)
21. Sudha, S., Vidhyalakshmi, M., Pavithra, K., Sangeetha, K., Swaathi, V.: An automatic classification method for environment: Friendly waste segregation using deep learning. In: 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 65–70. IEEE (2016)
22. Aral, R.A., Keskin, S¸.R., Kaya, M., Hacı¨omero˘glu, M.: Classification of trashnet dataset based on deep learning models. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 2058–2062. IEEE (2018)
23. Awe, O., Mengistu, R., Sreedhar, V.: Smart trash net: waste localization and classification. arXiv preprint (2017)
24. Yang, M., Thung, G.: Classification of trash for recyclability status. In: CS229 Project Report, vol. 2016 (2016)
25. Sousa, J., Rebelo, A., Cardoso, J.S.: Automation of waste sorting with deep learning. In: 2019 XV Workshop de Vis˜ao Computational (WVC), pp. 43–48. IEEE (2019)
26. Tzutalin: Labelimg. [Online]. Available: https://github.com/tzutalin/labelImg
27. Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H.: Cspnet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
28. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
29. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)

Design and Development of Road Surface Condition Monitoring System

Md. Imran Hossain, Mohammad Shafat Al Saif, Md. Rezaul Islam Biswas, Md. Seyam Mia, Abir Ahmed, and Md. Saniat Rahman Zishan

Abstract With an aging population, the demand of transportation is increasing. As a result, the road safety has become a critical issue. The road condition may differ from one place to another. Early detection of road condition can prevent number of accidents and casualties. In this paper, a system is proposed to reduce difficulties of the rider. The system can exert the unknown crippled point, potholes of any road by the acceleration of a vehicle whenever it passes away. Besides the position tracking and monitoring system, it will locate the respective location automatically and update current condition in Google Map. This location tracking information will help the upcoming rider to get the overview of the road beforehand.

Keywords Road monitoring · Safety alert · Accident reduction · Potholes detection

1 Introduction

Nowadays, the number of personal vehicles is increasing in Bangladesh because of the popularity of ride sharing services. People can save their time by using motor bike ride sharing service instead of public transport. As the number of vehicles increasing, road crashes have also risen markedly in last few years. As compared to other vehicles, motorcycle rider deaths number is alarming. Because of the road condition, immature driver and lack of attention on the road surface, number of accidents are increasing.

Md. I. Hossain · M. S. Al Saif · Md. R. I. Biswas · Md. S. Mia
Department of Electrical and Electronics Engineering, Faculty of Engineering, American International University-Bangladesh (AIUB), Dhaka, Bangladesh

A. Ahmed (✉) · Md. S. R. Zishan
Department of Computer Engineering, Faculty of Engineering, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
e-mail: abir.ahmed@aiub.edu

Md. S. R. Zishan
e-mail: saniat@aiub.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_29

In order to reduce road accident, there are many things that one needs to consider such as vehicles, breaking time and mostly road condition. After analyzing the reason that mostly responsible for accident, road condition detection is chosen to be solved in this paper. Accidents could be reduced if the rider is aware of the road condition before riding. It is necessary to find an efficient and low cost solution to find the abnormal surface in the road. One of the solution can be vibration acceleration, which is generated vibration if a vehicle passes through any abnormal surface [1]. This abnormal surfaces are also depending of the road capacity, traffic effect and other natural disaster based on situation [2]. To consider all this parameters, advance technology like machine learning can be used but it will make the system more complex. There are complex solutions available like calculating the road condition based on those parameters using advance neural network techniques [3] or instead of vibrational sensor, laser sensors can be used [4]. One of the techniques to collect data from surface is smart phone [5]. Data can be collected from smartphone with magnetometer and accelerometer sensors, then different algorithms, e.g., random forest (RF), support vector machine (SVM) will help to analyze the abnormal surface in the road [6]. This kind of advance techniques is required to quickly recognize the anomalies in the road and reduce the errors given by the system [7]. Road accidents sometimes increase during rainy season due to slippery road. So, detecting wetness of the road using planar capacitive sensors and updating the rider beforehand can also reduce accidents [8]. Accident can be reduced by accruing knowledge of tire-road. To detect the tire-road, friction co-efficient and forces need expensive sensors [9]. But in Bangladeshi condition, it will not help because of high cost. So, more cheap solution needs to be find out. Different microcontroller based solutions given by many researchers. This type of devices first collect data using accelerometers and then analyze the data to find out the anomalies. After finding out the location, it updates it server to take precautions for the next time [10, 11]. In 2011, the road condition was measured and composed by accelerometers, distance measuring instruments and graphic displays. It used the PRI and IRI analyzer with the pavement roughness level classifications [12]. In 2012, Pieto Bucci et al. propose a method to initialize the road condition, which can improved a road detection system in terms of mobile application GPS and accelerometer [13]. A monitoring system with linking the global monitoring data, real-time camera data, vibrating sensors, light meters and accelerometer was designed by Taylor et al. [14]. Measurement was done to particulate article exposures along routes with the road users. The idea of mobile sensing detection and reporting the surface visibility of roads was developed by MIT Computer Science and Artificial Intelligence Laboratory. They used the P2 architectures and accelerometer. The micro-electro-mechanical sensors (MEMS) will recognize the road surfaces based on measuring the acceleration force due to gravitational force “g” in three axes. By analyzing the continuous data flow from the sensor, the crippled position will be detected automatically via a microcontroller. Patricia et al. developed a system of identifying existing and emerging ITS technologies for both vehicles and motorcycles [15]. Another research from Kindai University Higashiosaka, Japan, authors developed an obstacle detection system for moving bicycle with the help of distance measuring sensors and they gave the graphical view of the avoidances [16]. This paper considers all of these aspect and proposed a model which will do simple calculation and easier to develop.

In this paper, the main objective is to give an indication of potholes, cracks and bumps, which affect driving comfort and on-road safety and the speed-breakers. It will ensure a safe ride by collecting the data for the next rider. This device will monitor the road continuously and find the potholes and bumps; then, it will store the position of those roughness. The sored data then will be uploaded in a modified Google Map, which will show the practical output to the user. This system will help the department of roads and highways to find out where to repair a road.

2 Architecture of the System

An automatic pothole detection device can help a rider in many ways. The device is fast enough to indicate the speed breaker and potholes and also locate the positions on a map so anyone can be aware of the condition of the road. In Fig. 1, the block diagram shows the whole process of the device, starting from sensing pothole to locating the positions on the map.

The whole device is operating based on a microcontroller device. It is using ARDUINO as the main microcontroller. To detect the potholes, a micro-electro-mechanical system technology-based sensor is used. The sensor will sense the potholes and sends a signal to locate the position. Arduino will take the location coordinates to locate and store the position data to a SD card. It will be uploaded to

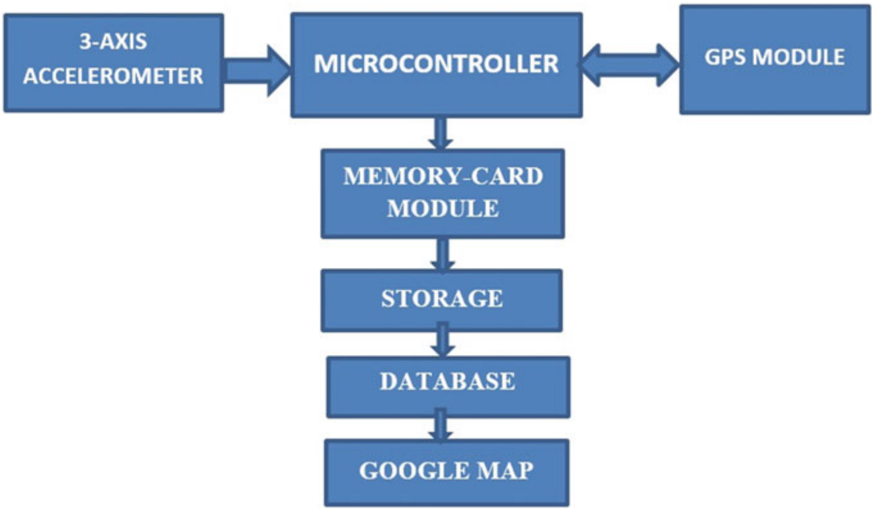


Fig. 1 Architecture of the proposed system

the map to see the condition of the road from anywhere in the world. A GPS module will give the exact position coordinate.

3 Simulated Result

Simulation of the system is divided into three parts according to the working principles: potholes detection, GPS locating and monitoring system. Each part individually simulated in Proteus 8.11 Software to verify the proposed model.

3.1 Detection of Potholes

Sensing the potholes is primary objective of this paper, the MPU-6050 sensor does the job of sensing, and the sensor will be placed in the device vertically. It works to sense the vibration and comparing it with the normal rating. After comparing the values, the Arduino will decide if it is a normal value or not. Figure 2 shows potholes detection simulation result in PROTEUS software.

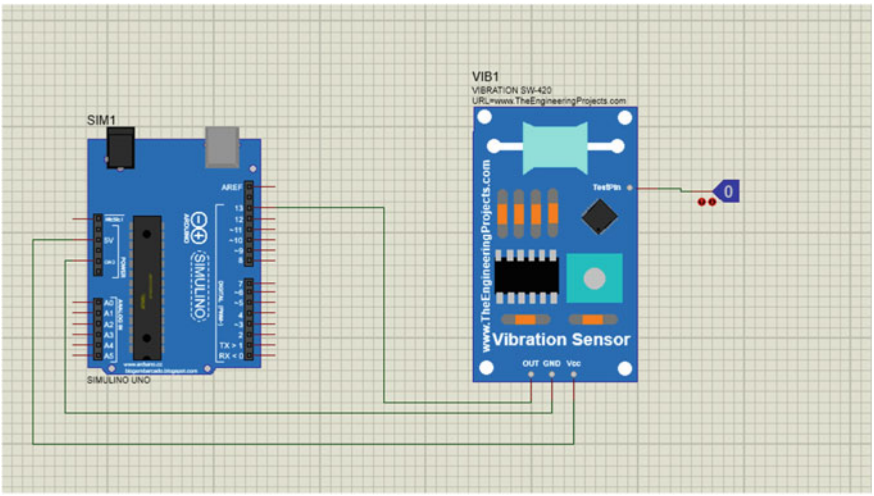


Fig. 2 Simulation result of potholes detection

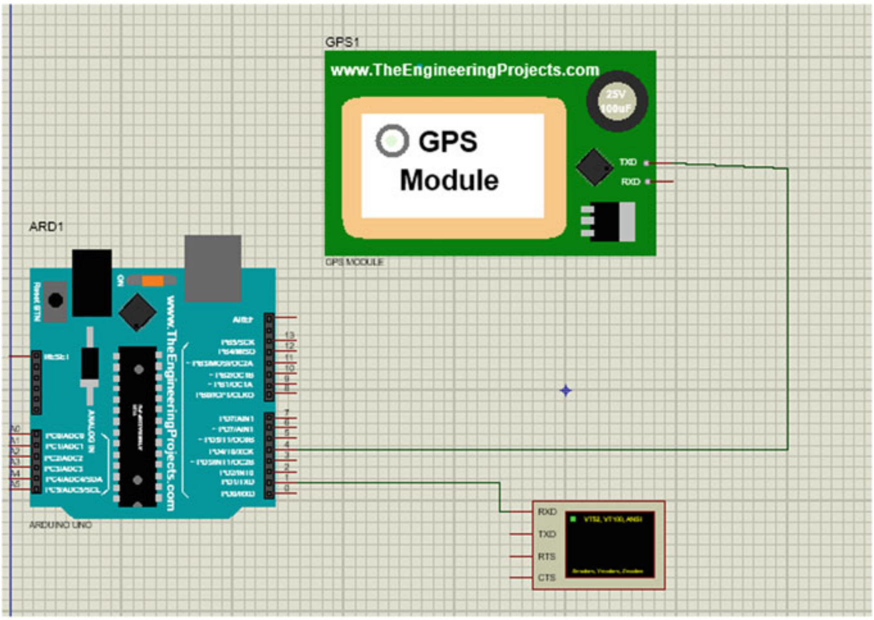


Fig. 3 Simulation result of GPS locating system

3.2 GPS Locating

The GPS locating system is designed to locate the potholes and give a coordinate value to the Arduino; then, the Arduino will store the coordinates to a SD card. The GPS module is inside the device, but the antenna is placed on the device to avoid connection with the satellite. The GPS send the coordinates continuously, but the Arduino will take only the coordinate when the sensor senses any potholes. Figure 3 shows the simulated result of GPS locating system.

3.3 Monitoring System

Figure 4 shows a monitoring system, which is designed to detect and find the location coordinates. The position of the coordinates is located on the map by the help of the online database called Google Fusion Table.

The sensor detects the potholes, then Arduino commands the GPS module to send the current location, when the Arduino gets the location coordinate, then it send command to SD card module to store the location coordinates to process further steps.

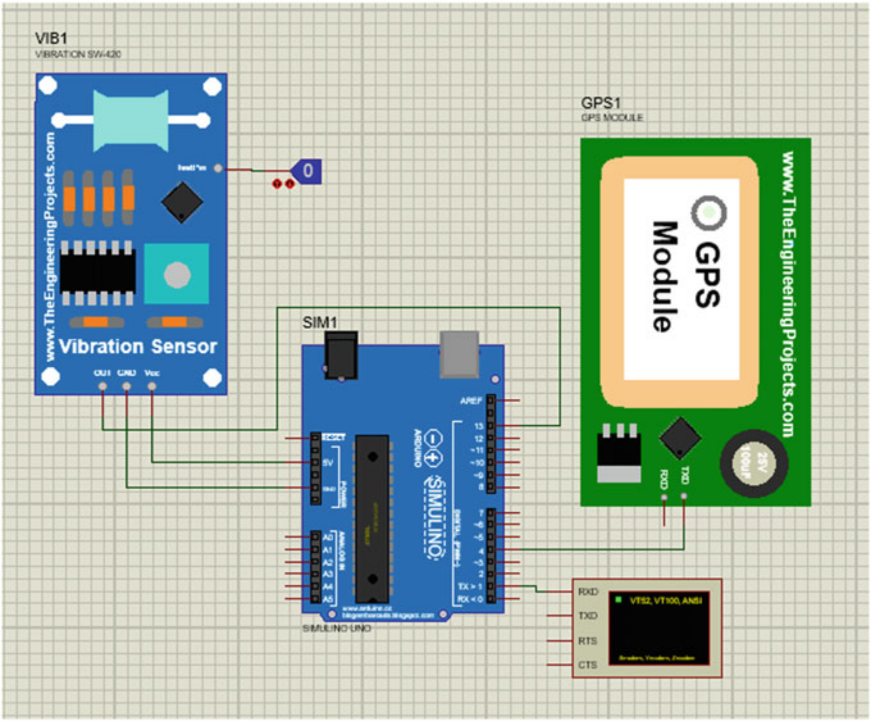


Fig. 4 Simulation result of monitoring system

4 Implementation

In this section, implementation of the proposed model is shown. The device size is small enough to setup in a bike easily. In the device, an acceleration sensor is used to detect the potholes and the bumps. A GPS module is used to locate the position, and the location coordinate is saved in a SD card to extract the location form the Google Map for the next rider. The detection of bumps and potholes and locating the position is done within a short period of time. All the coordinates saved in SD card are uploaded manually to the Google Map. The entire system is power up by a single DC battery, and all the process is controlled by an open source microcontroller named Arduino UNO.

All the sensors, modules and the microcontroller are combined in a box made of plywood. An additional power input port is available in the device in case of backup power supply. A high frequency antenna is attached on the top of the device, which can keep the connection of GPS to the satellite.

4.1 Implementing Pothole Detection

The roughness detection is done by a 3 axis ARDUINO accelerometer sensor (MPU-6050/GY-521). This sensor works by comparing the angle of inclination. After placing the sensor on a bike, the value of g for all 3 axes is noted for the normal condition. It is compared with the values that coming from the sensor at the time of riding. The analog data pins of the sensor are connected to the ARDUINO analog pins SCL to the A5 and SDA to the A4 pin as in this sensor database to obtain the values.

4.2 Locating the Position of the Bumps

A GPS module (neo 6 m v1) is used with a high frequency GPS antenna (1575.42 MHz) to locate the positions of the bumps or potholes. The GPS locates the position (latitude, longitude) when the sensor finds any roughness on the road. Rx. The communication with the GPS is done by the RX pin of the sensor in connected to the TX pin of the ARDUINO and TX of the sensor to the RX of the ARDUINO.

4.3 Storing Coordinates

The module communicates with ARDUINO through Serial Peripheral Interface (SPI) to store the data at SD card module. The SPI pins of the module CS, MOSI, MISO and SCK are connected to 10, 11, 12 and 13, respectively. The module can store data within short time. The procedure to insert the memory to the module is much easier. All the coordinates that coming from the GPS is stored in a text file in the SD card. In the text file, the coordinate will be saved the latitude first then the longitude. Both the values will be in a row, which is separated by a single coma. The next upcoming coordinates will be stored in next rows similarly.

A designed prototype is implemented after the successfully integration of every part of the model. A 5 mm PVC board is used to build the casing for the whole device. The charger port and the on off switch on the outside of the device which is shown in Fig. 5.

5 Result Analysis

The prototype tested in different situation and applied different level vibration. After analyzing the sensor values, calibration level for the sensor is set. The GPS module

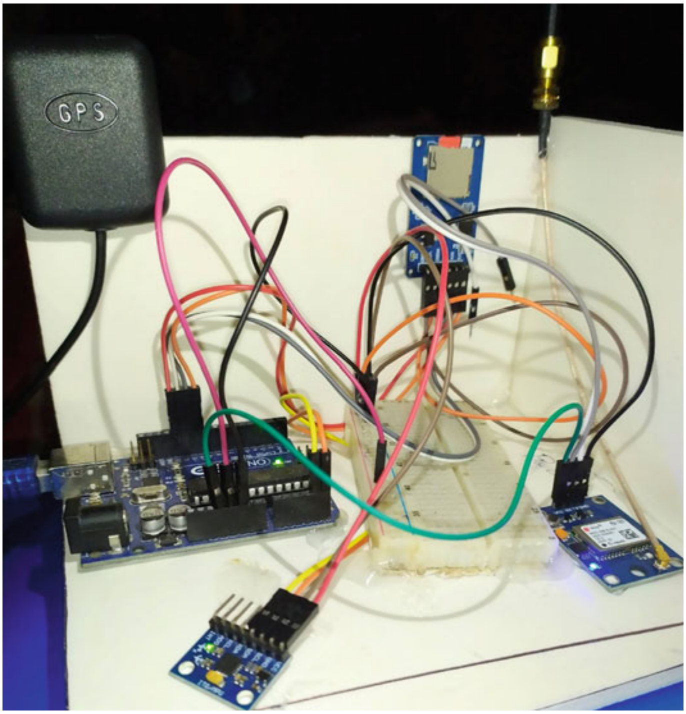


Fig. 5 Design prototype of the proposed model

is tested in different position indoor and outdoor. Variation in the result is found for different position.

5.1 Pothole Detection System

In this part of the work, the sensor is connected to the microcontroller to see the output. The sensor value is dependent on the many parameters like the position of the sensor, level of speed, amplitude of the vibration. It is observed that all the output of the sensor is responded to different situation and condition. Further, calculation is done from collected data to find out the appropriate value to set for the scenario. Then, calibration is done with the determined value by placing the sensor in different positions. At the normal position, the sensor value is around 16,000, in a normal vibration the value fluctuated from 1200–18,000. So, the value sets in the range between “900–20,000”. ARDUINO collects all value from the sensor and processed the data, then cross check with the range of set value. Thus, pothole is detected using the sensors.

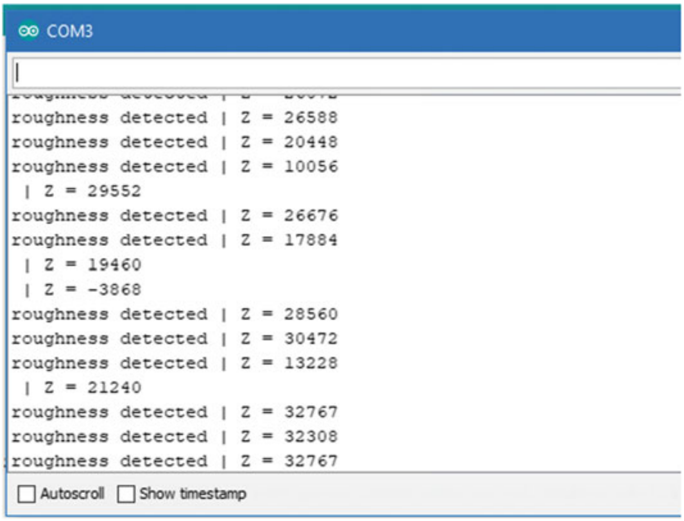


Fig. 6 Sensor data

Figure 6 is showing the sensor values for some bumpy point on the road, and the sensor gives a value along with the Z axis + 16,384 at a sensitivity of 2 g at normal condition.

5.2 Storing Value to SD Card

The values from the GPS module are saved on a SD card. In this process, a SD card module is used for the storing process. ARDUINO decide which value of the coordinates is to be stored, then it will send the value to the SD card module. The module will save the data to a text file that is created in the SD card previously. The values are separated by a single comma for every time that is coming. Figure 7 is showing the stored data in the SD card. The data stored every time in when the sensor getting any potholes or bumps.

5.3 Locating on the Map

After saving the data, the next step will be locating the data on map. In this section, the process to locate data is discussed. The text file that is created to store the data, which will be uploaded. Then, it uploads the file to a web site named Google Fusion Tables. This will locate the coordinates automatically to Google Map that will be the final output of this paper.

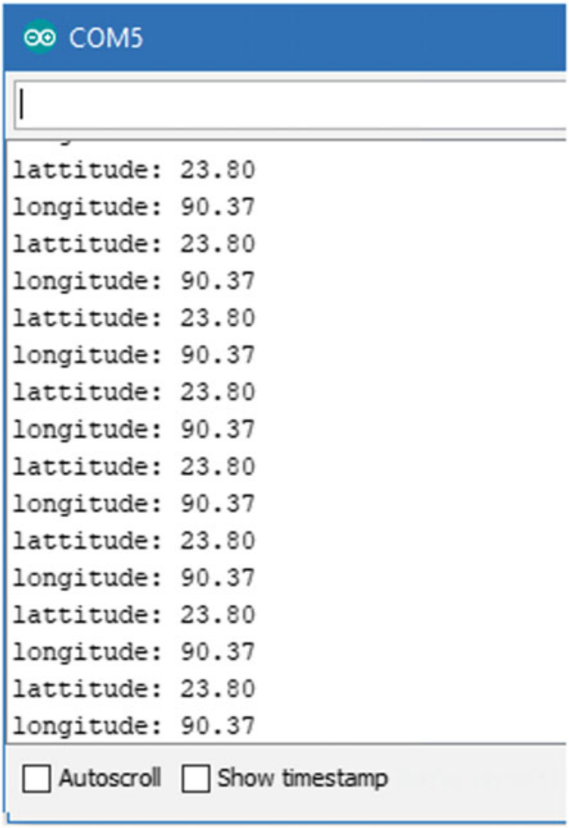


Fig. 7 GPS coordinates on seral monitor

Table 1 shows the server data after uploading the stored coordinates file on the Google Fusion Table. The latitude and the longitude that are separated by a comma and in separate columns in the file. Figure 8 is showing the positions on the Google Map from the Google Fusion Table. The potholes and the bumps located on the map by the red dots.

6 Discussion

In current situation, the road accident is a major problem in Bangladesh due to rapid increase of vehicles. Most of the time road condition is responsible for the accidents. Rider who is not accustomed with the road condition have greater risk to have an accident. Still in Bangladesh, riders are not used to with digital system or device from which a rider can know the road condition before starting the ride. Road surface monitoring system will reduce the problem, and the authority can take required steps for repairing the roads. The time and the cost of survey will also be

Table 1 Location tracked on a map

Detection No.	Latitude	Longitude
1	23.827750	90.418470
2	23.827815	90.419004
3	23.827693	90.418064
4	23.827781	90.418718
5	23.828012	90.419249
6	23.827781	90.417902
7	23.827720	90.418227
8	23.827726	90.418257
9	23.827735	90.418263
10	23.827782	90.418703
11	23.827787	90.418707

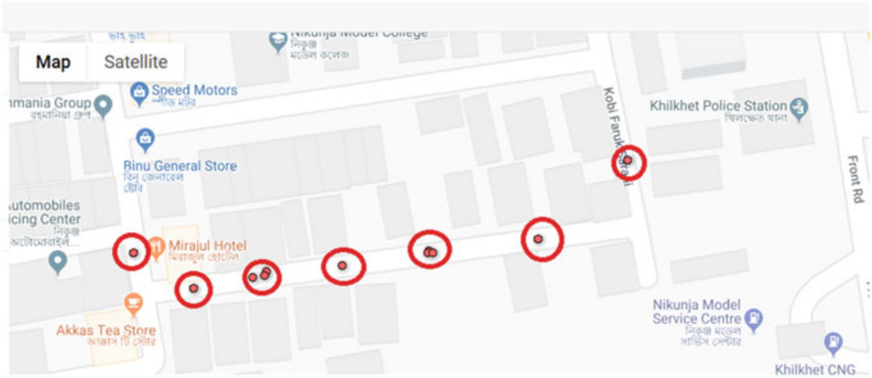


Fig. 8 Position tracking on the map

reduced for the developed system. The conditions of the roads are monitored, and the locations of bumps and the potholes are located on the Google Map, which could show the condition of the road to any user.

The cost of the device is only 3300 BDT, which is very affordable. The device size is small so it can be easily mounted on the bike. The system does not require any skill of the rider to detect the situation. The system is self-propelling. Thus, it can ensure both safety and comfortability of the rider while using new technology.

7 Conclusion

The main purpose of this paper is to introduce a platform where people riding bikes can quickly find information of the road condition before riding on that road and those who ride in the future will get information about where there are holes, cracks and barriers. Proposed system can be implemented to reduce the number of bike accidents on that road and to save their lives from the bike accident. This model will reduce the uncertainty of the road condition by updating data in Google Map in real time.

In the future, an auto balancing system can be developed for the biker on rough roads. A system can be developed, which use the rider's uploaded data automatically and suggest accordingly. Data accuracy can be improved by measuring the potholes condition measurement like deepness or size of it using machine learning techniques.

References

1. Du, R., Qiu, G., Gao, K., Lin, H., Liu, L.: Abnormal road surface recognition based on smartphone acceleration sensor. Sensors **20**(2), 451 (2020)
2. Akgul, M., Akburak, S., Hüseyin Yurtseven, A.O., Akay, H.K., Cigizoglu, M.D., Ozturk, T., Eksi, M.: Potential impacts of weather and traffic conditions on road surface performance in terms of forest operations continuity. Appl. Ecolo. Environ. Res. **17**(2), 2533–2550 (2019)
3. Xia, W., Chen, Z., Zhang, Y.-Z., Liu, J.: An approach for road material identification by dual-stage convolutional networks. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 7153–7156. IEEE (2018)
4. Li, W., Burrow, M., Li, Z.: Automatic road condition assessment by using point laser sensor. In: 2018 IEEE SENSORS, pp. 1–4. IEEE (2018)
5. Cong, J.L., Wang, Y., Yang, C.P.: Data preprocessing method of vehicle vibration acceleration by smartphone. J. Data Acquis. Process **34**, 349–357 (2019)
6. Dey, M.R., Satapathy, U., Bhanse, P., Mohanta, B.K., Jena, D.: MagTrack: detecting road surface condition using smartphone sensors and machine learning. In: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), pp. 2485–2489. IEEE (2019)
7. Salau, H.B., Onumanyi, A.J., Aibinu, A.M., Onwuka, E.N., Dukiya, J.J., Ohize, H.: A survey of accelerometer-based techniques for road anomalies detection and characterization. Int. J. Eng. Sci. Appl. **3**(1), 8–20 (2019)
8. Döring, J., Tharmakularajah, L., Happel, J., Krieger, K.-L.: A novel approach for road surface wetness detection with planar capacitive sensors. J. Sens. Sensor Syst. **8**(1), 57–66 (2019)
9. Cheng, S., Mei, M.-M., Ran, X., Li, L., Zhao, L.: Adaptive unified monitoring system design for tire-road information. J. Dyn. Syst. Measur. Control **141**(7) (2019)
10. Ambrož, M., Hudomalj, U., Marinšek, A., Kamnik, R.: Raspberry pi-based low-cost connected device for assessing road surface friction. Electronics **8**(3), 341 (2019)
11. Bidgoli, M.A., Golroo, A., Nadjar H.S., Rashidabad, A.G., Ganji, M.R.: Road roughness measurement using a cost-effective sensor-based monitoring system. Autom. Construct. **104**, 140–152 (2019)
12. Chen, K., Lu, M., Fan, X., Wei, M., Wu, J.: Road condition monitoring using on-board three-axis accelerometer and GPS sensor. In: 2011 6th International ICST Conference on Communications and Networking in China (CHINACOM), pp. 1032–1037. IEEE (2011)
13. Astarita, V., Caruso, M.V., Danieli, G., Festa, D.C., Giofrè, V.P., Iuele, T., Vaiana, R.: A mobile application for road surface quality control: UNIquALroad. Procedia Soc. Behav. Sci. **54**, 1135–1144 (2012)

14. Taylor, M., Fairfield, C.A. (2010). IntelliBike: condition monitoring of our cycling infrastructure. In: Bicycle and Motorcycle Dynamics Symposium on the Dynamics and Control of Single Track Vehicles, 20th-22nd October, Delft, The Netherlands (2010)
15. Eriksson, J., Girod, L., Hull, B., Newton, R., Madden, S., Balakrishnan, H.: The pothole patrol: using a mobile sensor network for road surface monitoring. In: Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services, pp. 29–39. ACM (2008)
16. Taniguchi, Y., Nishii, K., Hisamatsu, H.: Evaluation of a bicycle-mounted ultrasonic distance sensor for monitoring road surface condition. In: 2015 7th International Conference on Computational Intelligence, Communication Systems and Networks, pp. 31–34. IEEE (2015)

Adaptation of Blockchain Technology in the Mobile Banking System in the Context of Bangladesh

Md. Mainul Islam Roni, Mohammod Abul Kashem, Sonia Akter, Most. Fahmida Akter, Muhammad Riaz Hasib Hossain, and Mohammad Asaduzzaman Chowdhury

Abstract The mobile banking system is now very interested in blockchain technology because of its security mechanisms, data availability, increased transaction processing speed, and decentralized a database. Users of mobile financial institutions (MFIs) can use a mobile device to deposit money, withdraw cash, and send and receive money from their accounts. Mobile banking allows users to access account information, payments, deposits, withdrawals, and transfers, as well as investments, ATM assistance, and content services. The distributed ledger technology (DLT), often known as a blockchain, is a tamper-proof digital ledger that can store data and distribute it among connected nodes. Many scholars have contributed to the blockchain technology and consensus mechanism in their earlier studies. The process of integrating consortium blockchain technology and the Proof of Authority (POA) consensus protocol into Bangladesh's mobile banking system is the subject of this research. A model of blockchain architecture with its classification and consensus protocol is also discussed. Moreover, the 51% security level of POA consensus protocol is justified. This initiative will have a huge impact on the financial industry, as well as improve the security of mobile banking.

Md. M. I. Roni (✉) · M. A. Kashem · S. Akter
Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur, Bangladesh
e-mail: mainulislamrony@gmail.com

M. A. Kashem
e-mail: drkashemll@duet.ac.bd

Most. F. Akter
Center for Higher Studies and Research (CHSR), University of Professionals, Mirpur, Cantonment, Dhaka 1216, Bangladesh

M. R. H. Hossain
Department of Computer Science and Engineering, IUBAT, Uttara, Dhaka 1230, Bangladesh

M. A. Chowdhury
Department of Mechanical Engineering, Dhaka University of Engineering and Technology, Gazipur, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_30

Keywords Consortium blockchain · Proof of authority · Proof of work · Decentralized database · Smart contract

1 Introduction

The banking sector is always investigating to find out the better transaction process which will be safer, quicker, and cheaper. To foster market development, the Bangladesh Bank issued permits for mobile banking in July 2011. Five banks initially grabbed at the possibility to begin active deployments, with three of the largest doing so right ones and others doing so in early 2012. Brac Bank (bkash) and Dutch Bangla Bank Limited had the fastest early expansion at the end of the first quarter of 2012 [1]. Mobile banking refers to a system that allows customers to conduct banking transactions using their cell phone or other mobile devices. M-banking or SMS banking is also known as mobile banking. Bangladesh currently has the world's eighth-largest population (Bureau, 2017), making it one of the regions rising economies. However, when it comes to digital infrastructure, the country still lags behind other South Asian Association for Regional Cooperation (SAARC) members. The digital difference between its neighboring countries bears witness to the same conclusion. The rapid growth of Internet users in Bangladesh, on the other hand, has created several chances for new businesses such as e-commerce, m-commerce, e-banking, and m-banking. Even though some banks have established mobile banking platforms, the service has failed to gain attraction [2].

Blockchain technology revolutionizes the way secure transactions and data are stored on a network of devices. Cryptography is used to secure the data saved. Although this technology provides revolutionary change and more convenient services, the security challenges surrounding crypto technology are a critical subject that we must be addressed [3]. Blockchain is a method of recording and storing digital data that is nearly impossible to access by unauthorized parties. The system cannot be changed, hacked, or cheated in any way. It is an ideal solution for financial transactions. Authentication protocols are critical in blockchain because they allow only authorized parties to access the network as well as identify unauthorized parties [4]. Nodes are in blockchain technology connected in peer-to-peer networks and ledgers are stored in nodes. Blocks of digital data consist of distributed ledger following a set of rules, and each block is periodically added to the ledger [5]. The transaction using digital log is replicated and spread over the entire network on the blockchain [6].

A consortium blockchain is like a hybrid blockchain in terms of combining benefits for both private and public blockchains. Compared to a public blockchain network, a consortium blockchain network is safer, more scalable, and more efficient. As a private and hybrid blockchain, it enables access restrictions. Multiple selected organizations or groups of nodes are accountable for consensus and validating block in consortium blockchain [7].

The consensus algorithm is critical for preserving block security chains and efficiency. The appropriate algorithm can dramatically increase a blockchain application's performance [8].

In the recent years, smart contracts have grown in popularity to automate the execution of financial rules. The banking industry is looking for a solution of reducing paperwork that provides smart contracts. The banking industry has seen a significant transformation because of the adoption of this technology.

The rest of the paper has been organized in the following manner. Section 2 discusses the model of blockchain architecture and the classification of blockchain technology. Section 3 shows the concept of smart contracts, the usage of smart contracts in the blockchain and the usage of smart contracts in the mobile banking industry. The consensus protocol and several types of consensus protocols are explained in Sect. 4. Threats and attacks on the mobile banking system are discussed in Sect. 5. A proposed integrated blockchain technology model for the mobile banking system is shown in Sect. 6. Section 7 provides the findings of results for including the node validation process of POA consensus in the mobile banking system, the difference of 51% attack between POA and POW. The advantage of integrating consortium blockchain and POA consensus including in the mobile banking system is also discussed. Moreover, the research challenges, the solution to overcome the challenges, and the future work of this research are included. Finally, Sect. 8 draws a conclusion in this paper.

2 Model of Blockchain Architecture

Blockchain or distributed ledger technology (DLT) is a system of recording information and stores digital data in a way that unauthorized persons are difficult to access [9]. A blockchain is a network of blocks that carry several of unique identifications with a secure and legitimate manner (peer-to-peer). To put it another way, blockchain is a network of computers connected to one another instead of a central server, which means that the whole network is decentralized. In consequence, this system is not feasible for changing, hacking, or cheating. For financial transactions, it is a perfect solution. Authentication protocols play an important role in blockchain because it allows only the authorized parties to access the blockchain network and unauthorized parties must be found out [10].

Nodes are in blockchain technology connected in peer-to-peer networks and ledgers are stored in nodes. Blocks of digital data consist of distributed ledger following a set of rules, and each block is periodically added to the ledger [11]. Blockchain duplicates and distributes digital ledgers of transactions throughout its whole network of computer systems [12]. The best strategy for improving decentralization, accountability, equality, and transparency on the Internet is through blockchain. A distributed database of records is a blockchain that can be a public ledger or private ledger of digital records [13].

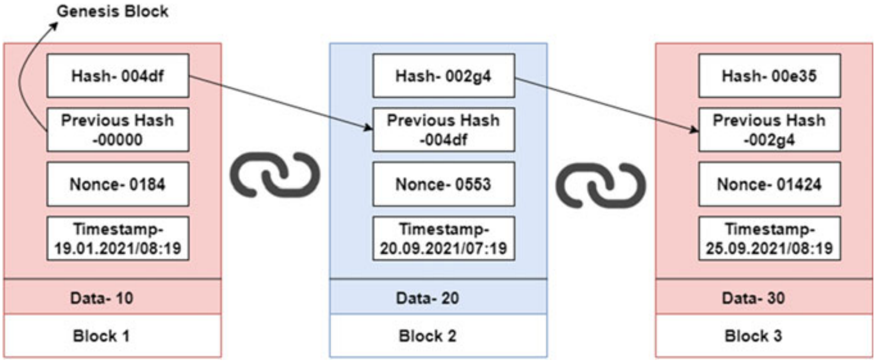


Fig. 1 Blockchain architecture

Figure 1 illustrates the blockchain architecture and the components of the blocks and connectivity of the blocks. Each ledger contains multiple blocks, and every block has the following components—main data, a hash of the former block, a hash of the present block, nonce value, and timestamp. A 32-bit whole number is referred to as a nonce. When a block is built, a random nonce is created, which is then used to generate a block header hash. The timestamp contains a small piece of data which is stored in each block as a single serial number, and its primary role is to determine the exact mining minute that has been approved by the blockchain network. Each block connects with every block with its previous hash. The first block is referred to as the genesis block because it has no previous block hash. Every hash value is encrypted by SHA-256 or SHA-512 algorithm. For that reason, if anyone wants to temper any block of data and update the chain with tempered data, the chain will refuse the tempered block.

2.1 Classification of Blockchain Technology

There are mainly four types of blockchain system. They are as follows (Fig. 2).

Public Blockchain or Permission Less Blockchain. Anyone on the Internet can join in the distributed ledger technology (DLT) and become an authorized node. The

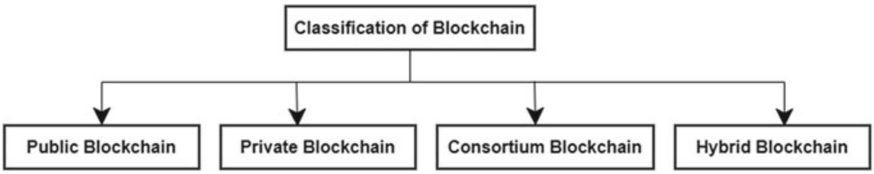


Fig. 2 Classification of blockchain technology

following things are possible using authorized nodes of public blockchain—access records, conduct mining activities, verify the transaction with complex computation, adding the transaction to the ledger [14]. Anyone in the network can verify transactions, uncover problems, and propose changes, thus no records or transactions can be modified. A public blockchain is independent and the network is slow. There is no restriction to access. The network can be hacked if 51% of the network gain by computational power [15].

Private Blockchain. Private blockchain is a closed and restricted network; only a single entity can use the network with permission and open for selective personnel who have permission to the network [16]. It may be small-scale organization or enterprise. There is a controlling organization that sets permission, authorization, and security for entity. The private blockchain can determine which node can display, add, or edit record in an organization. Private blockchain makes a faster network because of restriction of user limits. However, there is a claim that private blockchain is not an actual blockchain because of the main theory of blockchain is decentralization.

Consortium Blockchain. A consortium blockchain is a permissioned blockchain that is managed by multiple organizations. A public blockchain can be accessed by anybody, whereas a private blockchain can only be accessed by one company or organization [17]. A consortium blockchain is like a private blockchain in that it allows numerous organizations or groups of nodes to participate, whereas a private blockchain only allows one organization or corporation to participate. It is more secure, scalable, and efficient than a public blockchain network. It offers access controls as the private and hybrid blockchain.

Figure 3 illustrates how consortium blockchain works. Multiple groups of nodes are selected to participate in the consortium. Validation rules or consensus protocols are set to find out the valid node. Those nodes, which maintain the validation rules or consensus protocols, are selected as valid nodes and are not maintained the validation rules or consensus protocol which are selected as malicious or suspicious nodes in this consortium blockchain. These malicious nodes are not part of the chain. Then, the transaction queue and blocks of data are validated. When the transactions are validated as verified transactions, then the transactions broadcast to all valid nodes [18].

Hybrid Blockchain. The term “hybrid blockchain” refers to a network that combines both private and public blockchains [19]. Sometimes, organization wants to play a role as a private when it is necessary and which data are restricted and can be viewed by an authorized person. Restricted data are permission-based. Nevertheless, certain data are public and without authorization. User identity is protected from other users for security reasons. In hybrid blockchain, confidential information is also verifiable when needed. It is done by a different type of contract such as a smart contract. In the hybrid blockchain, privacy is protected but communication with a third party allows. Transactions offer better scalability and cheap and fast. Hybrid blockchain is not transparent completely because records can be shielded [20].

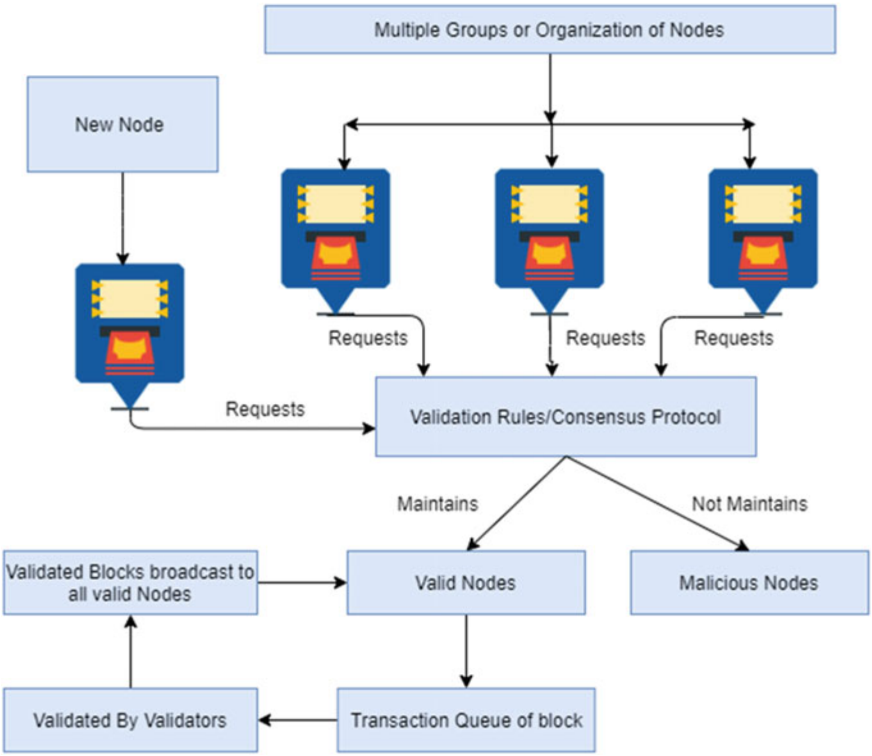


Fig. 3 How consortium blockchain works

3 Smart Contract

To create the automated execution of the business agreement, a smart contract creates a vital role in the recent times. It is the collection of code that is stocked in the blockchain and executed when the requirements are satisfied [21].

Smart contracts have a variety of advantages. They are resistant to tampering, self-executing, and self-verifying. They are also transforming the face of the banking business through error-free insurance claim processing, easy peer-to-peer transactions, faster KYC processes, transparent auditing, and other initiatives.

3.1 Why Smart Contracts Used in Blockchain

Smart contracts consist of a set of computer instructions that execute when specific conditions are fulfilled and are stocked on a blockchain. All parties are informed by automating execution of smart contracts and eliminating the wasted time and middleman work [22]. Smart contracts can also automate a workflow, initiating the next phase whenever certain criteria are met.

To carry out trustworthy transactions and agreements among disparate and anonymous participants, smart contracts do not require a centralized authority, a legal system, or an external enforcement mechanism.

3.2 The Usages of Smart Contract in the Mobile Banking Industry

Smart contracts may help a traditional mobile banking system in a variety of ways. The usages of the smart contract for the mobile banking industry are listed below-

Security. Smart contracts are a very secure mechanism for banking transactions because the smart contract overcomes the issue of dependency in the digital environment [23]. All transaction records in smart contracts are encrypted, making them exceedingly difficult to hack. Furthermore, because of each item on a distributed ledger is related to the preceding and following entries, hackers must modify the whole chain to change a single record.

Accuracy, speed, and efficiency. The contract is immediately executed when a condition is met. There is no paperwork to deal with because smart contracts are digital and automated, and no time is wasted rectifying errors that might arise when filling out paperwork manually [24].

Autonomy and Savings. The smart contract can contribute in many ways to mainstream banking. It is known that the banking sector has most of the activities depend on a third-party contractual agreement when the banking solution is purchased. In this situation, the smart contract can make trust and enforceable conditions with interested third parties for any case of a data breach [25]. It also makes other opportunities for a bank like insurance claims facility, almost real-time remittance service, and alternative auditing service and know your customer (KYC) facility by performing error inspection, routing, and a transfer fund to the handler if everything is matched with algorithm and originate proper. Moreover, smart contract also helps in trade payment, where the funds are transported one time the amounts of trade payments are deliberate accurately. Thus, smart contract can minimize the operational costs as well as administrative costs.

Transparency and trust. There is no third party involved in smart contracts, and encrypted records of transactions are exchanged between participants, making the transactions trustworthy.

4 Consensus Protocol

The banking sector, which handles millions of dollars in daily transfers from one part of the world to another, is one area where blockchain has had an influence. Consensus algorithms are a type of protocol that allows humans and machines to collaborate in a distributed, decentralized environment [26]. They contribute to keeping the network safe by verifying transactions. Before the blockchain is built, a consensus procedure is usually established. It is the heart of a blockchain network, specifies a technique for determining whether a transaction is valid or not. It offers a way to check and confirm what information should be put to a blockchain's record. Because blockchain networks often lack a centralized authority to determine who is correct or incorrect, all nodes on the network must agree on the network's state while adhering to set rules or protocols.

4.1 Types of Consensus Protocol

Many consensus mechanisms exist, including Proof of Work (POW), Proof of Stack (POS), and Proof of Authority (POA), among others [27].

Proof of Work (POW). Proof of Work (POW) is a consensus mechanism that is used to confirm a transaction and add a new block to the chain. Minors (a group of people) compete against each other in this procedure to complete the network transaction. Mining refers to the process of competing against one another. In each round of consensus, POW chooses one node or miner to produce a new block based on processing power. The participating nodes or miners must solve a cryptographic problem to win the puzzle. The node or miner that solves the problem first has the right to make a new block [28]. POW is a resource-intensive and power-hungry protocol. In comparison with more efficient protocols, this approach wastes a lot of computer power and electricity to solve cryptographic challenges [29].

Proof of Stack (POS). Unlike in Proof of Work, where a miner is paid to resolve mathematical problems and create new blocks, the creator of a new block is selected in a deterministic way based on its wealth, also known as stake [30]. This implies that the POS method does not have a block reward. Therefore, the miners collect transaction fees. The POS method has its own set of disadvantages, and actual application is exceedingly challenging.

Proof of Authority (POA). POA is a method for verifying and choosing nodes who are unquestionably trustworthy and willing to accept responsibility for the blockchain's security [31]. In POA, there is no technical competition between valid nodes as compared to the Proof of Work mechanism, which is commonly referred to as "mining". In this consensus, valid nodes approve the blocks of the transaction by a voting process. In this voting procedure (N/2) + 1, the most legitimate nodes must vote for validating new blocks to add them to the blockchain [32]. It means that when 51% of nodes put the vote for a block to validate, then this block will be integrated into the blockchain and broadcast to all valid nodes otherwise, this block will be identified as malicious data or suspicious data and it will be rejected to add inside the blockchain. By this process, the blocks store in the individual node, and the database will be decentralized and distributed database. To become a valid node, there will be some conditions [33].

1. A valid node must ensure their real identities.
2. A valid node needs to gain a reputation.
3. The selecting process of a valid node should be equal for all nodes.

This consensus process operates with very little computing power and, consequently, with very little electricity. To ensure that nodes are permanently connected, the POA algorithm does not need to solve problems. Because the POA only requires a small number of actors, the network can afford to update the blockchain more often by shortening the time between blocks (block time) and processing more transactions (block size) for near-zero processing costs (transaction fees) as well as it is possible to predict how long it will take for new blocks to be generated. This point changes for various consensus such as POW and POS.

5 Various Types of Threats and Attacks on Mobile Banking System

The mobile threat landscape is constantly changing, and both institutions and consumers are concerned about the dangers. Among today's rising worries are the following:

Mobile Malware. Today malware viruses are particularly designed for the mobile banking system. Some possible categories of attacks are [34].

1. Keylogger: Keylogger is a malicious attack that is used for stealing a user's log and sensitive information.
2. Spyware: Attacker can track a user's information from a smartphone by attacking spyware malicious program.
3. Trojan: The Trojan attack is the most familiar attack on the mobile banking system. It infected the network system connecting the router with the smartphone.

Usage of untrusted third-party applications or unsecured Wi-Fi. Sensitive information for a person or the entire system network can be observed by installing untrustworthy mobile applications or utilizing unprotected Wi-Fi.

Tampering Bank Data. By executing middleman and replay attacks in the communication media or network system, it is possible to alter the original data [35].

Phishing. Users lose access to their mobile devices when they open and click on links in SMS and e-mail messages. Mobile usage has created a slew of weaknesses, which fraudsters are keen to exploit.

Salami Attack. A salami attack is most associated with an automated alteration of financial systems and data.

6 Proposed Model of Integration Blockchain Technology in the Mobile Banking System

There are mainly four types of blockchain architecture, and many consensus protocols exist, but for the mobile banking industry, consortium blockchain and POA consensus protocol are the most preferable ones. Consortium blockchain architecture provides the facilities to add multiple groups of nodes inside an organization, and POA consensus protocol helps to find out the valid nodes and ensures the validity of the nodes.

Figure 4 shows the integration process of consortium blockchain and POA in the mobile banking system.

1. A user sends an authenticated transaction request to the acquiring bank.
2. A block is created to prepare a transaction.
3. Multiple groups of nodes are set to check the validation of the block using POA consensus protocols. Assume the number of valid nodes is N.
4. Making a voting process among valid nodes. All valid nodes must give a vote for each block of the transaction.

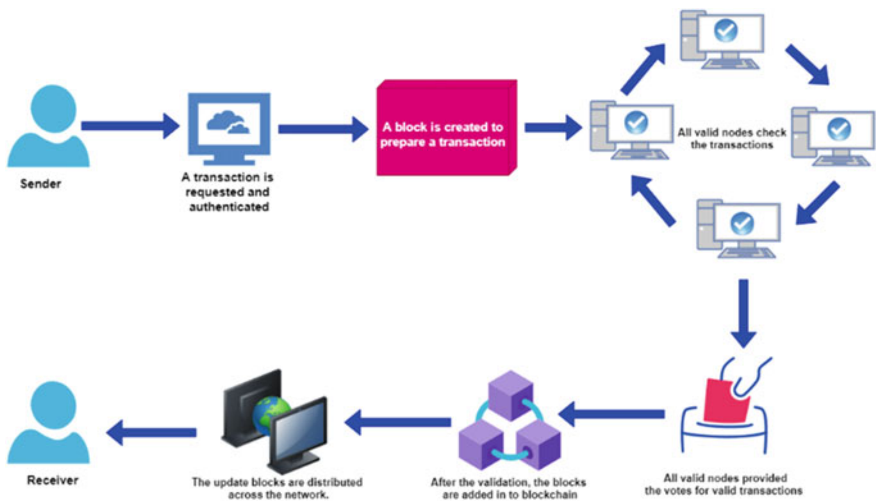


Fig. 4 Integration process of consortium blockchain and POA in the mobile banking system

5. If the block of a transaction gets maximum numbers of such as (N/2) + 1, then this block will add to the blockchain. Otherwise, this block will be marked as a suspicious or malicious block.
6. When a block is successfully added to the blockchain, the block will be updated into all distributed networks. After that, the receiver will get the money from the issuing bank.

7 Result and Discussion

Proof of Authority (POA) consensus protocol verifies the validity of the connected nodes and the requested nodes. Valid nodes are solely involved in the POA consensus, as shown in Fig. 5. Each node must prove its valid identity and acquire a reputation to be considered a genuine node. With the current six legitimate nodes, two more nodes are added to the mix. For the additional two nodes, actual identities (such as Mac addresses, IP addresses, and so on) are necessary, as well as the ability to build a reputation. About 51% clearance of the eight nodes is necessary to transact any of the nodes.

In Table 1, the existing valid nodes (EVN) and new valid nodes (NVN) are explained as the calculation of the minimum voting of POA consensus protocols. N = EVN + NVN calculates the total number of nodes, whereas (N/2) + 1 calculates the minimum number of votes. Initially, several existing valid nodes are essential during the integration of blockchain technology with the POA consensus in the mobile banking system of Bangladesh. According to Table 1, a bank has twenty existing valid nodes for blockchain integration with POA mechanism at the beginning, thereafter ten new valid nodes are appealed to take part in the vote. In this case, minimum sixteen votes or appreciations are required among the thirty nodes for the first transaction. After completing the first transaction, the number of existing valid nodes is increased, and twenty new valid nodes have been added to the system. As a result, minimum votes are raised into twenty-six. During the third transaction, no new node is invited for joining due to that, the minimum number of votes is stable on the system. In the transaction four, fifteen new valid nodes are offered, so that, the number of minimum votes has been increased into thirty-four. In this way, the minimum number of votes will be seventy-one for the transaction of fifteen whose valid total will be one hundred forty following the addition of eleven nodes.

The number of minimum votes increases when additional legitimate nodes are added to the blockchain, as seen in Fig. 6. The percentage rate is constant here, although the number of votes and valid nodes varies. When a result, as more legitimate nodes join, the number of minimum votes grows. Furthermore, new legitimate nodes will be addressed to the blockchain frequently. Under the circumstances, hacker needs to have 51% control over legitimate nodes to induce any transaction blockages, which is quite difficult.

To entice the blockchain technology, both POA and POW consensus require a 51% attack. An attacker must control 51% of the network of legitimate nodes in

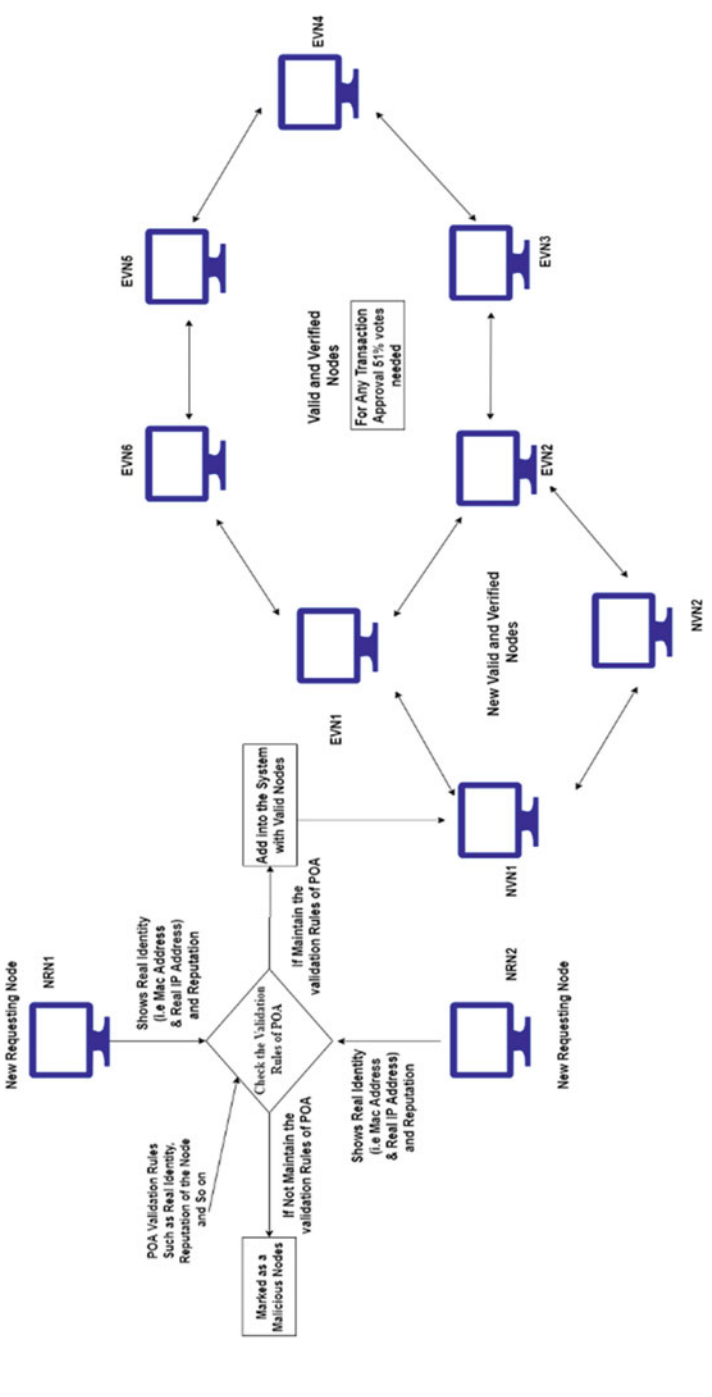
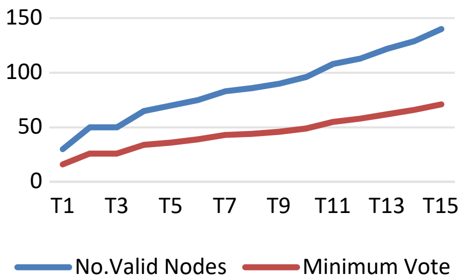


Fig. 5 Structure for node validation (EVN refers to an existence valid node, NVN stands for a new valid node, and NRN denotes a new requesting node)

Table 1 POA security level (T denotes transaction number, EVN expresses existing valid nodes, NVN marks new valid nodes, and N indicates a total number of valid nodes)

T No.	EVN	NVN	N = EVN + NVN	Minimum vote, $(N/2) + 1$
01	20	10	30	16
02	30	20	50	26
03	50	0	50	26
04	50	15	65	34
05	65	05	70	36
06	70	05	75	39
07	75	08	83	43
08	83	03	86	44
09	86	04	90	46
10	90	06	96	49
11	96	12	108	55
12	108	05	113	58
13	113	09	122	62
14	122	07	129	66
15	129	11	140	71

Fig. 6 Validation nodes versus minimum vote



POA consensus, but in POW, an attacker must control 51% of the network computational power [36]. It is feasible to lure the data in POW if an attacker obtains 51% of the processing capacity of network nodes. Researchers recently demonstrated that a selfish miner with 33% mining power may obtain 50% mining power [37]. Computational power, on the other hand, has no bearing in POA.

7.1 Advantage of Integration Consortium Blockchain and POA in the Mobile Banking System

Although blockchain technology is a new concept in Bangladesh, the acceptance of blockchain technology is growing rapidly around the world particularly in mobile banking and financial sectors for its advanced functionalities. The benefits of integrating consortium blockchain and POA protocol are given below.

1. Intermediary: Consortium blockchain introduces the decentralization and distributed system. It will help to reduce intermediary costs.
2. Authentication and identification: POA ensures the authentication of nodes and checks the identification by the reputation of the nodes.
3. Reduce Manual Agreement Processing: Smart contract reduces the manual agreement and paperwork.
4. Less Power Consumption: POA does not need any computational power. In POA consensus protocol, the node does not need any extra electricity for its operation. Thus, it saves power consumption.
5. Maintain High Security: POA maintains the high security of the block. About 51% nodes need access to tempt the block, and the minimum number of votes is changed by adding nodes. So, it is quite impossible to tempt a single block of data.
6. Availability of Data: In consortium blockchain, data are always available within valid nodes.
7. Integration of New Node: The consortium blockchain and POA have a fantastic feature that allows to install a new legitimate node.

7.2 Research Challenge

The major challenge accompanying consortium blockchain is a lack of awareness of technology in Bangladesh, especially in the mobile banking system, and a well-known understanding of blockchain technology works in banking sector. Consortium blockchain saves transaction cost and time but initially, it needs a high capital cost. Moreover, trust between actors is difficult to establish in an audit and coordination situation. In addition, a lack of research paper is also a research challenge.

8 Conclusion

The banking industry realizes that it requires better exploit technology for securing mobile banking sector in Bangladesh. Blockchain technology can be one of the possible answers. In this regard, the design of consortium blockchain allows to build on a set of nodes for an organization, and the POA consensus mechanism contributes to determine the genuine node as well as their authenticity. It is noted that faster block times, a better transaction throughput, and a higher level of security are the advantages of the consortium blockchain and POA consensus methods. The focus of this study is on the methods for ensuring data security on the mobile banking business in Bangladesh. There will be a huge revolution in data security if this technology is used in the mobile banking system. In addition to, blockchain technology can make the potential changes underlying the credit information system, loan management system, payment clearing system, and digital verification of user's using smart contract and KYC. Although, blockchain is new a technology in Bangladesh, this paper provides a comprehensive overview of blockchain technology that is critical to the implementation of the mobile banking sector in Bangladesh. However, further research is required on blockchain technology to handle data redundancy and big data management.

References

1. Islam, M.M., Hossain, Md.E.: An investigation of consumers' acceptance of mobile banking in Bangladesh. Int. J. Innov. Digit. Econ. **6**, 16–32 (2015). https://doi.org/10.4018/ijide.2015070102
2. Islam, M.S., Karia, N., Mohamed Soliman, M.S., et al.: Adoption of mobile banking in Bangladesh: a conceptual framework. Rev. Soc. Sci. **2**, 01 (2017). https://doi.org/10.18533/rss.v2i8.109
3. Ghumman, F.: Distributed ledger technology: secuirty threats in cryptography. SSRN Electron. J. (2019). https://doi.org/10.2139/ssrn.3393539
4. Madhuri, S., Nagalakshmi, V.: A Survey on Authentication protocols for Block-Chain Technologies and Cross-Chain Transactions 9 (2021)
5. Khan, P.W., Byun, Y.-C., Park, N.: A data verification system for CCTV surveillance cameras using blockchain technology in smart cities. Electronics **9**, 484 (2020). https://doi.org/10.3390/electronics9030484
6. B.R., D.P., KB, B.: Monetization of private data using blockchain ensuring control and ownership of the user data. Int. J. Progres. Res. Sci. Eng. **2**, 17–19 (2021)
7. Al-Saqaf, W., Seidler, N.: Blockchain technology for social impact: opportunities and challenges ahead. J. Cyber Policy **2**, 338–354 (2017). https://doi.org/10.1080/23738871.2017.1400084
8. Puthal, D., Malik, N., Mohanty, S., et al.: Everything you wanted to know about the blockchain: its promise, components, processes, and problems. IEEE Consum. Electron. Mag. **7**, 6–14 (2018). https://doi.org/10.1109/MCE.2018.2816299
9. Du, M., Xiaofeng, M., Zhe, Z.: et al.: A review on consensus algorithm of blockchain, pp, 2567–2572 (2017). https://doi.org/10.1109/SMC.2017.8123011
10. Cui, Z., Xue, F., Zhang, S., et al.: A hybrid blockchain-based identity authentication scheme for multi-WSN. IEEE Trans. Services Comput. 1–1 (2020) https://doi.org/10.1109/TSC.2020.2964537
11. Lewis, R.: Blockchain and financial market innovation. Econ. Perspect. **13** (2017)
12. Helo, P., Hao, Y.: Blockchains in operations and supply chains: a model and reference implementation. Comput. Ind. Eng. **136**, 242–251 (2019). https://doi.org/10.1016/j.cie.2019.07.023
13. Niranjanamurthy, M., Nithya, B.N., Jagannatha, S.: Analysis of blockchain technology: PROS, CONS and SWOT. Cluster Comput. **22**, 14743–14757 (2019). https://doi.org/10.1007/s10586-018-2387-5

14. Gatteschi, V., Lamberti, F., Demartini, C., et al.: Blockchain and smart contracts for insurance: is the technology mature enough? Future Internet **10**, 20 (2018). https://doi.org/10.3390/fi10020020
15. Park, J.H., Park, J.H.: Blockchain security in cloud computing: use cases, challenges, and solutions. Symmetry **9**, 164 (2017). https://doi.org/10.3390/sym9080164
16. Gordon, W., Catalini, C.: Blockchain technology for healthcare: facilitating the transition to patient-driven interoperability. Comput. Struct. Biotechnol. J. **16** (2018). https://doi.org/10.1016/j.csbj.2018.06.003
17. Li, Z., Kang, J., Ye, D., et al.: Consortium Blockchain for Secure Energy Trading in Industrial Internet of Things (2017)
18. Zheng, Z., Xie, S., Dai, H., et al.: An overview of blockchain technology: architecture, consensus, and future trends. In: 2017 IEEE International Congress on Big Data (BigData Congress), pp. 557–564 (2017). https://doi.org/10.1109/BigDataCongress.2017.85
19. Sagirlar, G., Carminati, B., Ferrari, E., et al.: Hybrid-IoT: Hybrid Blockchain Architecture for Internet of Things—PoW Sub-blockchains. arXiv:180403903 [cs] (2018)
20. Thwin, T.T., Vasupongayya, S.: Blockchain based secret-data sharing model for personal health record system. In: 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA), pp. 196–201 (2018). https://doi.org/10.1109/ICAICTA.2018.8541296
21. Luu, L., Chu, D.-H., Olickel, H., et al.: Making Smart Contracts Smarter, pp. 254–269 (2016). https://doi.org/10.1145/2976749.2978309
22. Wright, A., De Filippi, P.: Decentralized blockchain technology and the rise of lex cryptographia. SSRN Electron. J. (2015). https://doi.org/10.2139/ssrn.2580664
23. McKinney, S., Landy, R., Wilka, R.: Smart contracts, blockchain, and the next frontier of transactional law. Washington J. Law, Technol. & Arts **13**, 313 (2018)
24. Peters, G., Panayi, E.: Understanding modern banking ledgers through blockchain technologies: future of transaction processing and smart contracts on the internet of money. SSRN Electron. J. (2015). https://doi.org/10.2139/ssrn.2692487
25. Giancaspro, M.: Is a 'smart contract' really a smart idea? Insights from a legal perspective. Comput. Law Secur. Rev. **33**, 825–835 (2017). https://doi.org/10.1016/j.clsr.2017.05.007
26. Salah, K., Rehman, M.H.U., Nizamuddin, N., Al-Fuqaha, A.: Blockchain for AI: review and open research Challenges. IEEE Access **7**, 10127–10149 (2019). https://doi.org/10.1109/ACCESS.2018.2890507
27. Ismail, L., Materwala, H.: A review of blockchain architecture and consensus protocols: use cases, challenges, and solutions. Symmetry **11**, 1198 (2019). https://doi.org/10.3390/sym11101198
28. Zhang, S., Lee, J.-H.: Analysis of the main consensus protocols of blockchain. ICT Express **6**, 93–97 (2020). https://doi.org/10.1016/j.icte.2019.08.001
29. Sayeed, S., Marco-Gisbert, H.: Assessing blockchain consensus and security mechanisms against the 51% attack. Appl. Sci. **9**, 1788 (2019). https://doi.org/10.3390/app9091788
30. Chaudhry, N., Yousaf, M.: Consensus Algorithms in Blockchain: Comparative Analysis, Challenges and Opportunities, pp 54–63 (2018)
31. Di Silvestre, M.L., Gallo, P., Guerrero, J., et al.: Blockchain for power systems: current trends and future applications. Renew. Sustain. Energy Rev. **119**, 109585 (2019). https://doi.org/10.1016/j.rser.2019.109585
32. Nguyen, T., Kim, K.: A survey about consensus algorithms used in Blockchain. J. Inf. Proc. Syst. **14**, 101–128 (2018). https://doi.org/10.3745/JIPS.01.0024
33. De Angelis, S., Aniello, L., Lombardi, F., et al.: PBFT vs proof-of-authority: applying the CAP theorem to permissioned blockchain (2017)
34. Wazid, M., Zeadally, S., Das, A.: Mobile banking: evolution and threats: malware threats and security solutions. IEEE Consum. Electron. Mag. (2019). https://doi.org/10.1109/MCE.2018.2881291
35. Becher, M., Freiling, F., Hoffmann, J., et al.: Mobile security catching up? Revealing the nuts and bolts of the security of mobile devices. In: 2011 IEEE Symposium on Security and Privacy (2011). https://doi.org/10.1109/SP.2011.29

36. Li, X., Jiang, P., Chen, T., et al.: A survey on the security of blockchain systems. Futur. Gener. Comput. Syst. **107**, 841–853 (2017). https://doi.org/10.1016/j.future.2017.08.020
37. Eyal, I., Sirer, E.G.: Majority is not enough: Bitcoin mining is vulnerable. Commun ACM **61**, 95–102 (2018). https://doi.org/10.1145/3212998

IoT-Based Smart Energy Metering System (SEMS) Considering Energy Consumption Monitoring, Load Profiling and Controlling in the Perspective of Bangladesh

Mehedi Hasan Jewel, Parag Kumar Paul, Abdullah Al-Mamun, and Md. Shahriar Ahmed Chowdhury

Abstract Electricity usage is directly connected to the socio-economic growth of a society. As a developing country, the increased demand for electricity supply in Bangladesh has increased the emphasis on the need for accurate and reliable energy measurement methods. The estimation of the electricity consumption from households, institutions and industries is a must for a government to profile the country's power capabilities and to plan for new installations if needed. From the perspective of the developing countries, mining authentic information about the ratio of the electricity consumption to its production is a matter of dilemma due to the conventional metering and billing system for the consumed electricity. In this paper, we have proposed a design which eradicates many of the drawbacks of the current system by integrating the energy measurement technology with IoT. Xtensa dual-core 32 bit microprocessor, Class 1 single phase energy measuring device, MODBUS, Wi-Fi, GSM/GPRS as communication protocols, MQTT as server, and MySQL as database system have been used in the proposed system. In this paper, real-time energy data monitoring, load control and emergency alert scheme have been demonstrated. Besides, the daily load-wise electricity consumption has been estimated throughout a week to analyze the fluctuations of three different loads considering their electricity consumption (kWh) and the balance credited (3.75BDT/kWh) on a daily basis. The output of this research has been found electrically, statically and economically very significant, which can be implemented as the next generation IoT based Smart Energy Metering System (SEMS) in the perspective of Bangladesh.

Keywords IoT · Energy meter · Load control · Server · Web application · Load profiling

M. H. Jewel · P. K. Paul · A. Al-Mamun · Md. S. A. Chowdhury (✉)
Centre for Energy Research, United International University, Dhaka 1212, Bangladesh
e-mail: shahriar.ac@gmail.com

P. K. Paul
e-mail: ppaul141084@bseee.uiu.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_31

1 Introduction

Electrical power has become part and parcel of the day to day lives of human beings. It has become a necessity and a very important factor in the progress and advancement of the society and the economy of a society. The demand for electricity has been on the rise all over the world and correspondingly, also in Bangladesh. It is estimated that by the end of the year 2021, the total electricity demand in Bangladesh will be around 20,000 Megawatt [1]. Research done by Bangladesh Power Development Board (BPDB) shows that since 1981, the GDP of Bangladesh has been on the rise in correspondence with the annual electricity consumption [2].

An energy meter is an electrical device which measures the energy consumption of any household, industry etc. In a traditional metering system, electromechanical devices were used for energy measurement. This energy measurement system was flawed as it could be easily tampered with and the meter readings could be manipulated. There have been numerous cases where the consumers were being over-billed by the authority, which is referred to as "ghost electricity bills" [3]. In order to eradicate the problems caused by the traditional metering technology, the Bangladesh Government introduced a digital prepaid metering system, in which consumers are able to use as much as electricity as they have paid for [4]. Each prepaid meter uses a SIM card and a prepaid card analogous to the SIM card. Due to the reliability of prepaid meters, the energy authorities of Bangladesh has planned to install 3.85 crore prepaid energy meters in Bangladesh in the fiscal year 2020–2021 [5]. Even though prepaid meters are more reliable and robust compared to the traditional postpaid meters, they still lacks some very important features such as independent, remote and real-time monitoring of the consumption of electricity, load control, online prepay mechanism etc. Moreover, there have been multiple complaints about overbilling the consumers in prepaid meters by the billing authority [6].

Taking all the issues of traditional meters and currently used prepaid meters into account, we have designed a Smart Energy Metering System (SEMS) that incorporates energy measurement with the Internet of Things (IoT). The term "IoT" refers to the technology that connects the sensors, processors and computers to the networking system. In this system, the energy consumption of the household machines and devices are not only measured, but are also uploaded to the cloud server via the MQTT protocol which can be accessed via a website or a mobile application for real-time monitoring. The user and the authority can access an individual household's daily/weekly/monthly consumption at any given time and analyze the load consumption data. The user can manually control individual loads by turning it on or off remotely or as a demand response mechanism in order to adjust with the generation profile. Moreover, an automatic online billing system is introduced where the consumers can pre pay the electricity bills using an online payment gateway. A web application and a mobile application have been developed for real-time monitoring, load control, payment etc. Another important feature in our system is low balance alert system where the consumers are alerted via text message in case of low electricity balance. An emergency loan system is also integrated into the system where any consumer can take a loan of 100 taka maximum in case of discontinuation of electricity due to low balance.

2 Literature Review and Analysis of Related Works

We analyze the research objectives and outcomes as well as the drawbacks of the researches done previously. The summary of the review can be seen in Table 1.

As seen from Table 1, the researches done previously in this category has various drawbacks. In our proposed system, we have tried to addressed the lacking of previous works.

3 Proposed Design and Implementation

3.1 System Architecture

The current energy consumption measurement and billing system in Bangladesh has limitations, such as the lack of a real-time monitoring system, no daily/weekly/monthly electricity consumption record, no load control mechanism and low balance alert mechanism. In our system we have overcome the limitations posed by the current system as well as the drawbacks and limitations of previous researches.

The complete architecture of the system can be seen in Fig. 1. As seen from the architecture, the IoT based networking is based on GSM/Wi-Fi technology. At the beginning of the system, the Intelligent Control Unit (ICU) initializes the GSM and Wi-Fi module to connect to the network. After a successful connection, the system establishes connection with the MQTT based cloud server. After initializing all the sensors and network connections, MODBUS communication protocol is used in the collection of meter readings from the energy measuring device and the meter data are sent to the connected cloud server for real-time monitoring. At each cycle the ICU checks for available balance of the user and when the balance reaches below 100 Taka, an emergency alert text message is sent to the user's phone via GSM module. An emergency loan system has also been introduced in the system where any consumer will be able to take a loan of up to 100 Taka in case of unwanted power outage due to low balance. The working flow-diagram of the proposed system is described by the flowchart in Fig. 2.

Table 1 Summary of previous researches and the corresponding drawbacks

Reference #	Proposed	Limitations
Mortuza et al. [7]	GSM based automatic energy meter reading and billing system	This proposed system fails to describe an IoT platform for remote monitoring of the consumption of electricity
Mahfuz et al. [8]	Smart energy meter and digital billing system for Bangladesh	This proposed system introduces an SMS alert-based communication system but fails in terms of remote monitoring and online payment gateway for instant billing
Redwanul Islam et al. [9]	An IoT based real-time low-cost smart energy meter monitoring system using android application	The developed system here has been included with an IoT platform which can be accessed by a website and/or mobile application. The drawback is it does not introduce a load control mechanism
Sahani et al. [10]	IoT based smart energy meter	The design architecture proposed by this research lacks an online payment gateway for payment and also an emergency alert system. The web application developed by the researcher is not user friendly and lacks a data storage system for future assessment
Ashna et al. [11]	GSM based automatic energy meter reading system with instant billing	The system developed in this research includes a user-friendly platform for real time consumption monitoring, instant billing system and emergency alert. The drawback is that it does not include a load control scheme
Hlaing et al. [12]	Implementation of WiFi-based single phase smart meter for Internet of Things	This research includes real-time monitoring and load profile analysis but fails to account for individual load control, emergency alert system and instant billing scheme
Chaudhari et al. [13]	Smart energy meter using Arduino and GSM	The proposed system uses Arduino as the microprocessor and GSM as the network connectivity technique, but fails to describe a real-time data monitoring platform, data storage system, automated instant billing scheme etc

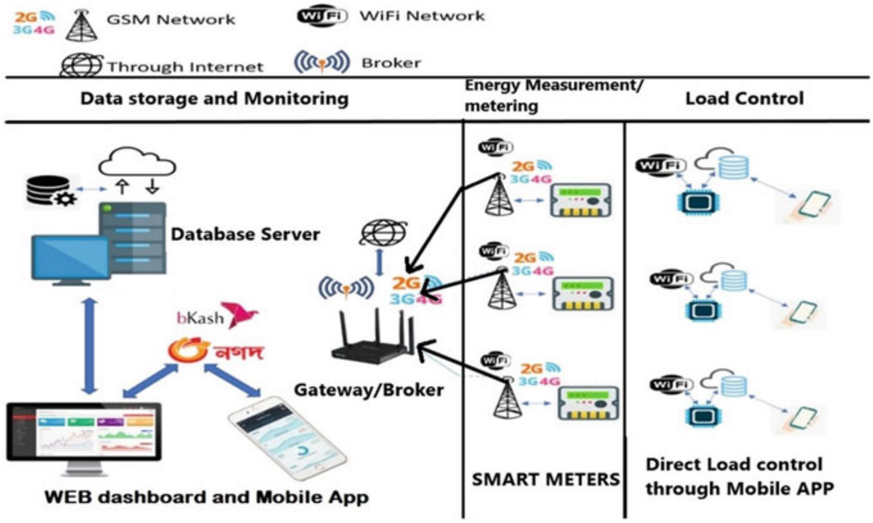


Fig. 1 Proposed architectural model of smart energy metering system

3.2 Hardware Description

Intelligent Control Unit (ICU). ICU or the intelligent control unit is the main processing unit of the system. This is the part that is responsible for the processing of the data gathered from the digital energy meter. In our system we have used an Xtensa dual-core 32 bit low power microprocessor that can operate at a rate of 240 MHz clock speed. It holds a memory of 320 kb of Ram and 448 kb of ROM. One of the main features of this processor is a built in Wi-Fi: 802.11 b/g/n connectivity.

Energy Measuring Device. The device is used in the measurement of real-time energy usage. In our designed system we have used Class 1/Class energy measuring device. The device is IEC 62,053–21/EN50470-1/3 international standard certified. This bi-directional energy measuring device is able to measure both active and reactive power of the electrical network. This device also supports RS485 MODBUS RTU interface for remote reading of the measured data [14].

Communication Protocol. In order to establish communication between the ICU and the energy meter, MODBUS serial communication protocol is used. As the energy meter that we have used in our system supports MODBUS RS-485 RTU, we have opted for RS-485 as the MODBUS type. RS-485 is a serial data transmission standard which is used widely in industrial implementations. One of the advantages of using the RS-485 communication protocol is that it is a duplex communication system where multiple devices on the same bus can create bi-directional communication [15].

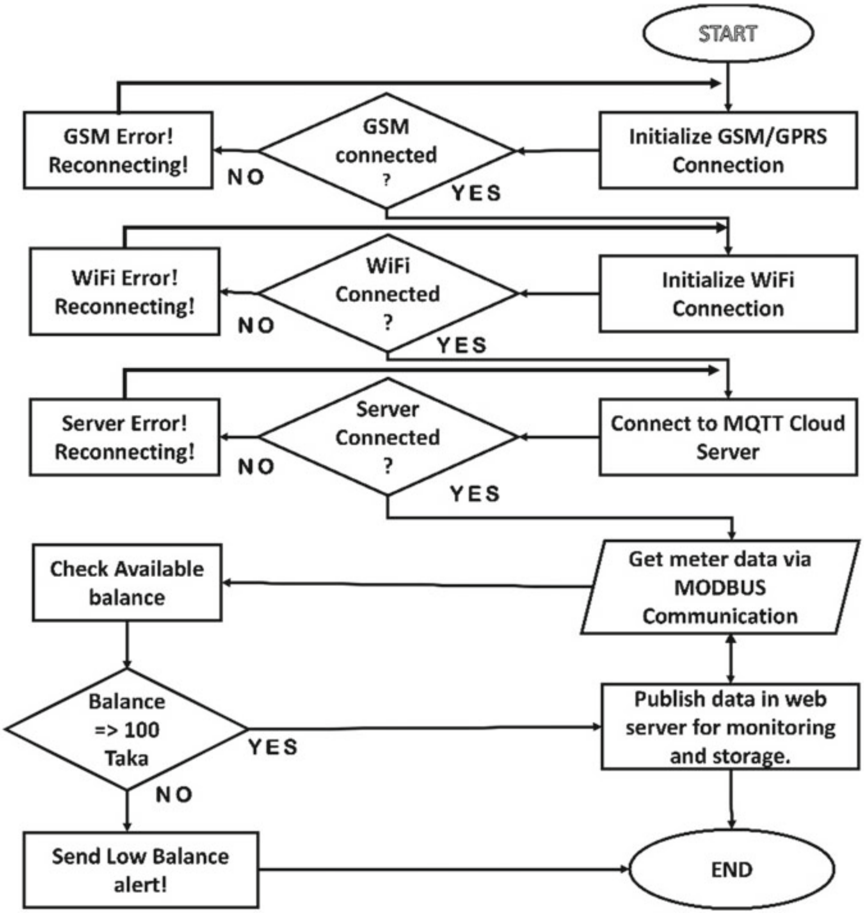


Fig. 2 Overall flow diagram of the working principle of proposed smart energy metering system

GSM/GPRS Communication. The GSM modem used in our system is a piece of hardware which includes multiple communication protocols within the system. The TTL output and the RS232 output are used in communication with the microcontroller and computers respectively. It can connect to any global GSM network with any 2G SIM card [16]. The objective of using a GSM/GPRS communication device are threefold.

- To connect to the internet/cloud server via GPRS in order to publish meter readings for real-time monitoring and storage.
- To send "Low Balance Alert" to the corresponding consumer.
- To receive instant loan command from the consumer in case emergency balance is needed to continue electricity supply.

3.3 *Software Tools*

MQTT Broker. An MQTT broker refers to a client-based server that receives messages from the clients and re-routes the messages to the destination. It is a many-to-many communication protocol for interchanging messages between multiple clients. The callback function of the MQTT server enables users to publish specific messages to a pre-determined topics that the device is subscribed to [17]. Due to its fast response and secure connectivity, we have used this protocol to transfer meter data to the cloud servers [18]. In our proposed system, all the data from the energy meter are published to an MQTT server via MQTT protocol.

Data Storage and Management. In order to store the meter data for future analysis and load profiling, a data storage system has been adopted. In our proposed system, we have used MySQL database management system. MySQL is a free and open-source software under General Public License. A python script runs behind the database management system (DBMS). This DBMS stores the crucial consumer information, daily/weekly/monthly consumption data etc.

3.4 *Design and Implementation of the System*

Initially, the system was designed and simulated using the Proteus Design Suit prior to the printed circuit board development. The design schematic of the system can be seen in Fig. 3.

4 Experimental Outputs and Results

The developed system was experimented on customized electrical circuit board with 3 different light bulbs simulating 3 different loads. The main control circuit was connected with the energy meter and the meter readings were processed prior to uploading them into the MQTT server. In Fig. 4, we can see the main circuit board and its operation.

4.1 *Real-Time Monitoring of Energy Consumption*

The energy data measured by the energy meter are collected and processed by the ICU and uploaded instantly to the web server via Wi-Fi/GSM module. The decrypted data received by the server and the web-dashboard displaying the received data can be seen in Fig. 5a and b. The JSON format of the received data is:

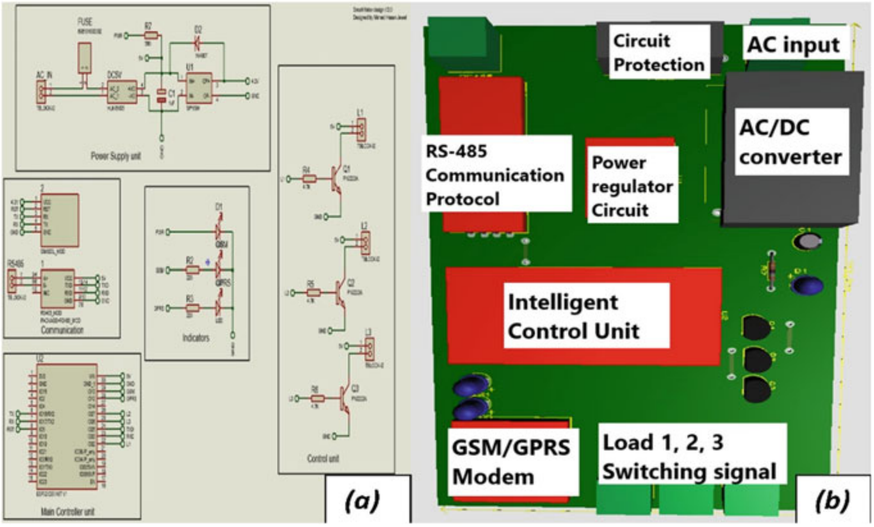


Fig. 3 Schematic (**a**) and 3-dimensional model (**b**) of the main control circuit

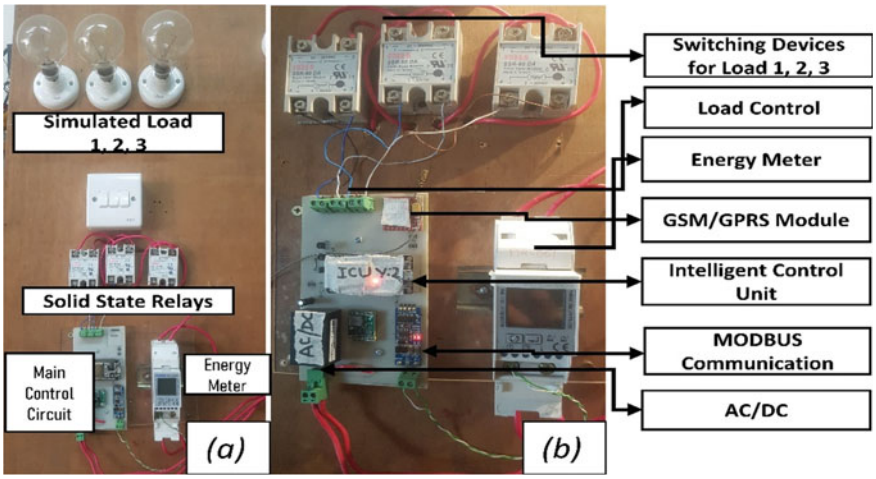


Fig. 4 Operational preview of developed smart energy metering system (SEMS) (**a**) and background model of main control circuit interfacing with the energy measuring device (**b**)

{“voltage”:“220.2”, “current”:“2.42”,"power”:“0.394”,“re”:“00.00”,“freq”:“50.12”, “en”:“10.68”, “pf”:“0.9”,“pr”:“0”}.

In order to access the website, each user is assigned a unique ID and password. The authorized person needs to login with the corresponding credentials before he/she can monitor the real-time readings. A mobile application was also developed which can be seen in Fig. 6.

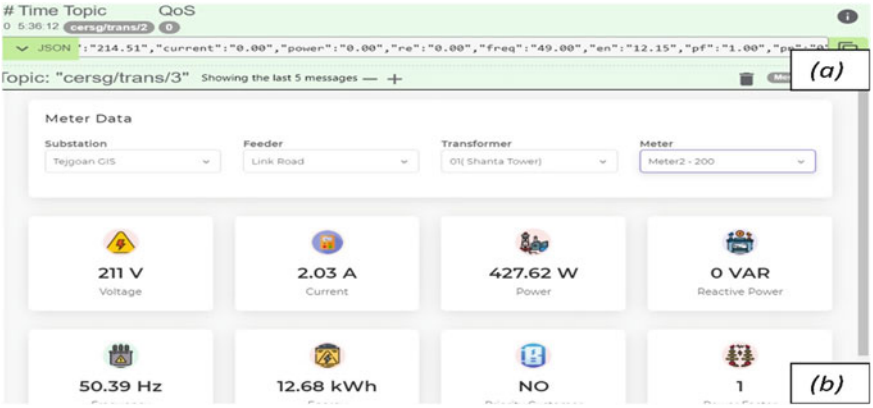


Fig. 5 Decrypted meter data format received by the server (**a**) Smart Energy Metering System (SEMS) dashboard (**b**)

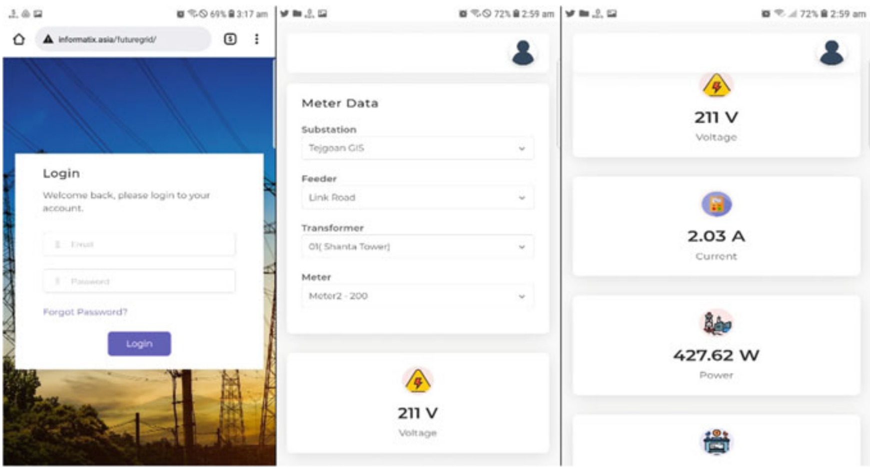


Fig. 6 Mobile application for real-time energy consumption monitoring with load control

4.2 *Electricity Consumption Data Analysis and Load Profiling*

In our proposed system, the hourly electricity consumption data of each connected load is stored in the database which can be accessed by authorized people. In this way, the faulty electricity billing and overbilling issue can be avoided. In order to simulate the data storage system of the proposed device, we connected various types of loads with load 1, 2 and load 3 as shown in Fig. 4. Load 1 was connected with heavy loads such as air-conditioner, refrigerator etc., load 2 was connected with various electrical household appliances which can be categorized as medium loads.

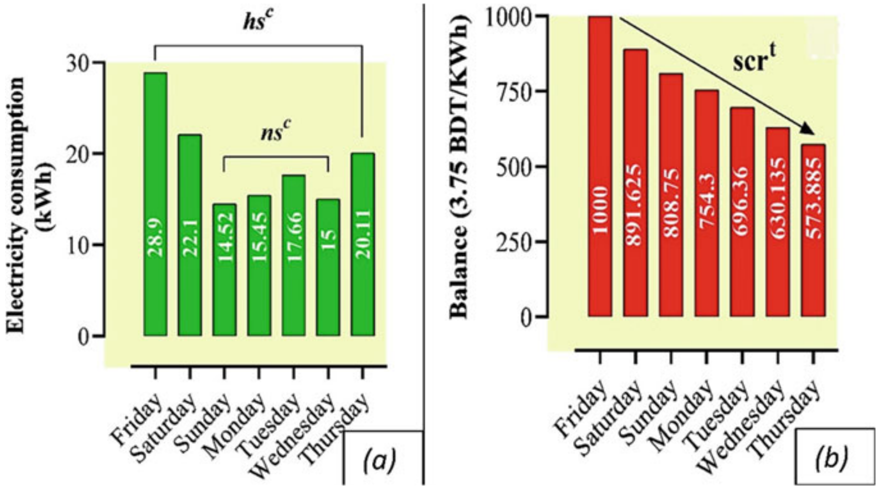


Fig. 7 Graphical representation of electricity consumption (**a**) and amount of credit expensed in accordance with the consumption (**b**) **Legends:** hs^c—high significant consumption, ns^c—non significant consumption, scr^t: significant cost reduction over time

For example, television, irons and load 3 was connected to lighter loads. After a week of measuring, monitoring and storing the daily consumption, the load data was analyzed as seen in Fig. 7.

The load profiles created from the available data from the SEMS website is presented in Fig. 7a. We can see that electricity consumption was maximum in Friday. 28.9 kWh of electricity was consumed that day. The minimum amount of electricity was consumed on Sunday with 14.52 kWh. On average, in Friday and Saturday, consumer's consumption of electricity is much higher as explained by hs^c. On the other hand, Sunday, Monday and Wednesday sees low energy consumption which is described by ns^c. From Fig. 7b, we can see the initial electricity balance of BDT 1000 is being adjusted each day in correspondence with the daily consumption. Here, the cost of electricity is 3.75 BDT/kWh (for lifeline consumers).

More in-depth data can be analyzed as seen in Fig. 8 where energy consumption of each connected load per day is graphically represented. On average, load 1 consumed the maximum electricity everyday which is as expected. Load 2 and load 3 consumes less electricity compared to load 1. The difference between the 3 load types was maximum during Sunday and Monday, described by s^f and the least during Tuesday described by ns^f.

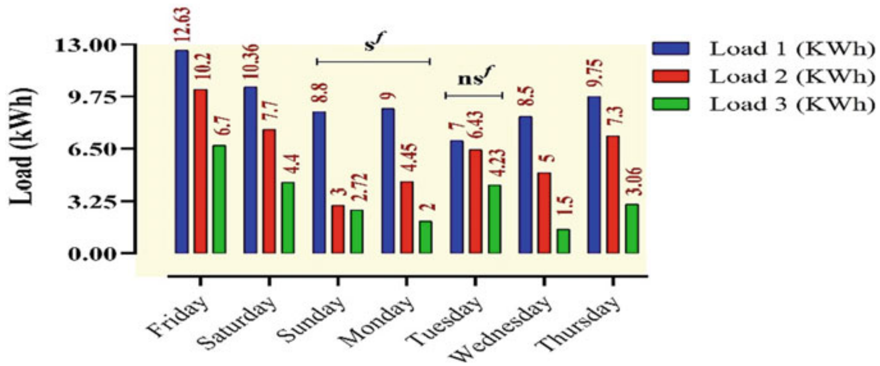


Fig. 8 Load-wise daily consumption of electricity **Legends:** s^f—significant fluctuation, ns^f—non-significant fluctuation

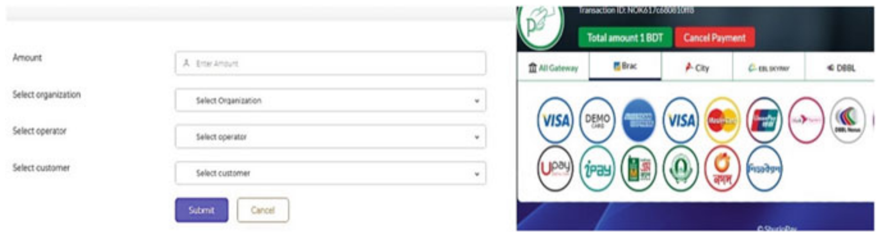


Fig. 9 Payment method of the online payment gateway

4.3 Prepaid Bill Payment System

An instant billing mechanism has been integrated into the Smart Energy Metering System. The consumers and users can login to the website or mobile app and select "Bill Payment" option to pre pay the electricity bill. The billing system overview from the website is presented on Fig. 9.

4.4 Emergency Alert and Loan Scheme

An emergency SMS or text alert system has been integrated into the proposed system in order to notify the consumers about low electricity balance. The algorithm for this alert mechanism is seen in the flow chart presented in Fig. 2. When the balance is less than 100 taka, a "Low Balance" alert is sent to the corresponding consumer via the GSM module. In special cases, when the consumer runs out of balance and electricity is cut-off by the authority, an emergency loan of 100 taka can be requested. The "Low Balance" alert received by the customer is seen in Fig. 10.

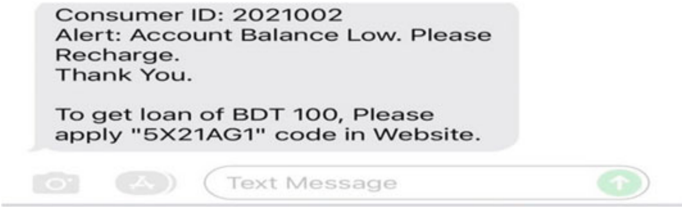


Fig. 10 Low balance alert in consumer's text message

In order to ensure reliability of the system, an alternative alert system has been incorporated where if the Smart Energy Metering System is disconnected from the server for more than 20 min, a "Device Failure" alert is sent to the billing authority from the server.

5 Discussion and Conclusion

The current prepaid metering technology used in Bangladesh has limitations. Some of the main limitations are lack of online monitoring system, no electricity consumption data storage system, load control, online billing scheme etc. Due to lack of proper monitoring, ghost billing or overbilling issues have been reported. In this paper we have presented a solution to the problems associated with the current energy metering technology in Bangladesh. We have incorporated IoT with the energy measurement technology and implemented many features that ensure reliability and robustness of the system. Our proposed system measures the energy parameters such as voltage, current, frequency, reactive power and sends the data over to the cloud server for real-time monitoring via MQTT protocol. The data are stored in the database which can be accessed anytime by the authorized person to evaluate the consumption and the bill in order to avoid any faulty billing. A load control technology has also been incorporated into the system where the user can control 3 different loads using the mobile app or the web browser in order to reduce electricity usage and save cost. The online prepaid billing system eases the life of consumers by saving valuable time and the emergency alert and loan system improves the overall experience of both the consumer and the authority. Integrating algorithm to detect electricity theft and automatic demand response system to reduce electricity loss and increase the reliability of the electrical network can be part of future researches.

Acknowledgements This work is done as part of the research project funded by Bangladesh Energy and Power Research Council (BEPRC), Power Division, Ministry of Power, Energy and Mineral Resources (MPEMR), Government of Bangladesh (Grant No: EPRC/58-2018-002-01).

References

1. Energy and Mineral Resources Division. https://mpemr.gov.bd/power/details/80. Last accessed 25 Oct 2021
2. Masuduzzaman, M.: Electricity consumption and economic growth in Bangladesh: co-integration and causality analysis. Glob. J. Manag. Bus. Res. **12**(11), 46–56 (2012)
3. Customers suffer as power company continue to overbill. https://www.dhakatribune.com/bangladesh/power-energy/2020/06/23/customers-suffer-as-power-companies-continue-to-overbill. Last accessed 26 Oct 2021
4. Jain, A., Bagree, M.: A prepaid meter using mobile communication. Int. J. Eng. Sci. Technol. **3**(3) (2011)
5. Over 3.85cr electricity consumers will get prepaid meters. https://www.tbsnews.net/bangladesh/energy/over-385cr-electricity-consumers-will-get-prepaid-meters-193222. Last accessed 20 Oct 2021
6. What led to ghost electricity bills. https://www.dhakatribune.com/bangladesh/2020/06/27/what-led-to-ghost-electricity-bills. Last accessed 27 Oct 2021
7. Mortuza, M.G., Ahmed, Z.U., Jaki, H., Ahmed, M.R., Uddin, M.J., Kabir, M.H.: A GSM based automatic energy meter reading and billing system. Euro. J. Eng. Technol. Res. **5**(6), 725–730 (2020)
8. Mahfuz, N., Nigar, M., Ulfat, N.: Smart energy meter and digital billing system for Bangladesh. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–4. IEEE (2020)
9. Islam, M.R., Sarker, S., Mazumder, M.S., Ranim, M.R.: An IoT based Real-time Low Cost Smart Energy Meter Monitoring System using Android Application. arXiv preprint arXiv:2001.10350 (2020)
10. Sahani, B., Ravi, T., Tamboli, A., Pisal, R.: IoT based smart energy meter. Int. Res. J. Eng. Technol. (IRJET) **4**(4), 96–102 (2017)
11. Ashna, K., George, S.N.: GSM based automatic energy meter reading system with instant billing. In: 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), pp. 65–72. IEEE (2013)
12. Hlaing, W., Thepphaeng, S., Nontaboot, V., Tangsunantham, N., Sangsuwan, T., Pira, C.: Implementation of WiFi-based single phase smart meter for Internet of Things (IoT). In: 2017 International Electrical Engineering Congress (iEECON), pp. 1–4. IEEE (2017)
13. Chaudhari, S., Rathod, P., Shaikh, A., Vora, D., Ahir, J.: Smart energy meter using Arduino and GSM. In: 2017 International Conference on Trends in Electronics and Informatics (ICEI), pp. 598–601. IEEE (2017)
14. Fitzgerald, D.E.: Development of an energy monitoring practice framework to profile energy use in residential buildings. Doctoral dissertation, Murdoch University (2021)
15. RS-485 Serial Interface Explained. https://www.cuidevices.com/blog/rs-485-serial-interface-explained. Last accessed 21 Oct 2021
16. Kanani, P., Padole, M.: Real-time location tracker for critical health patient using arduino, GPS Neo6m and GSM Sim800L in health care. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 242–249. IEEE (2020)
17. Mishra, B.: Performance evaluation of MQTT broker servers. In: International Conference on Computational Science and Its Applications, pp. 599–609. Springer, Cham (2018)
18. MQTT Security Fundamentals. https://www.hivemq.com/blog/mqtt-security-fundamentals-tls-ssl/. Last accessed 18 Oct 2021

Use of Genetic Algorithm and Finite-Difference Time-Domain Calculations to Optimize "Plasmonic" Thin-Film Solar Cell Performance

Abrar Jawad Haque, Mustafa Mohammad Shaky, Sabrina Nurhan Hasan, Tawseef Ahmed Khan, M. Shamim Kaiser, and Mustafa Habib Chowdhury

Abstract A methodology based on the use of finite-difference time-domain calculations and Genetic Algorithm is proposed that predicts aluminum nanoparticles (NPs) can be used for improving the opto-electronic performance of thin-film solar cells. Aluminum nanoparticles were coated with a thin silica shell layer that resulted in shifting the absorption properties of aluminum nanoparticles to longer wavelengths and aiding the chemical isolation of the highly reactive aluminum nanoparticle core. Silicon thin-film solar cells were then modified with various sizes of aluminum nanoparticles with varied shell thicknesses that were placed on top of the silicon substrate, embedded inside it, and placed in a "sandwich" configuration, with one particle on top of the substrate and another embedded inside, respectively. The results showed that Al-silica core–shell nanoparticles in a "sandwich" configuration demonstrated the most improved opto-electronic performance of solar cells when compared to the other configurations studied. These promising results can open the door for future collaborative use of Genetic Algorithm and FDTD calculations to design high sensitivity photovoltaic devices.

Keywords Genetic algorithm · FDTD · Core–shell · Nanoparticle · Solar cell · Aluminum · Plasmonics · Surface plasmon resonance

1 Introduction

In this era of scarce natural resources and the constant search for sustainable energy source, renewable energy, especially solar energy is showing tremendous potential due to its abundance and sophisticated technological availability. A major focus has been created globally to harness this enormous energy accounting for around

A. J. Haque · M. M. Shaky · S. N. Hasan · T. A. Khan · M. H. Chowdhury (✉)
Department of Electrical and Electronic Engineering, Independent University, Bangladesh, Dhaka, Bangladesh
e-mail: mchowdhury@iub.edu.bd

M. Shamim Kaiser
Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_32

3,400,000 EJ reaching the Earth each year [1]. A large barrier in implementation of solar energy is production of solar cells, which is of very high cost. Among the production cost, 40% accounts for wafer fabrication of the solar cell [2]. Thin-film solar cells are one type of solar cells of only a few micrometers thick and comprises few light absorbing materials, thus they are low in efficiency with low electrical current generation capability. Several methods are under investigation to improve the performance of these thin-film solar cells among which the use of plasmonic nanostructures is showing good potentiality. Some metallic nanostructures exhibit a phenomenon called localized surface resonance (LSPR) [3, 4]. This phenomenon can be utilized to increase the opto-electronic performance of thin-film solar cells [5, 6]. Among different materials, aluminum (Al) is known to have good optical and plasmonic characteristics. However, aluminum is a reactive metal and in contact with air tends to react vigorously forming aluminum oxide coating on its body. When illuminated with light, different metals demonstrate different extinction spectra [7]. Aluminum has its peak extinction spectra in the UV region of visible light spectrum, which is its another drawback. Coating of aluminum core with a few nanometers thick silica shell is one method to shift the resonant peak spectra of the aluminum towards the optical absorption region of amorphous silicon, along with chemical isolation of the nanoparticle from the surrounding material, keeping the aluminum nanoparticle core non-reactive.

Thus, for this study, aluminum nanoparticle core is coated with silica (SiO_2) shell for maximum interaction with the TFSC. Extinction spectra of homogenous (i.e., with no shell layer) Al nanoparticles (NP) were compared with Al NP with various thickness of silica shell, to observe the shift in resonant wavelength. Later, three different configurations were chosen to observe the opto-electronic performance of the thin-film solar cells with Al nanoparticle. Al NP on top of the silicon substrate, Al NP embedded inside the substrate and Al NPs in sandwich configuration (Fig. 1). The performance of the three configurations was assessed individually by optical and electrical parameters with various diameter of the core, various thickness of the silica shell and different depths of the NP when embedded inside the substrate as described below.

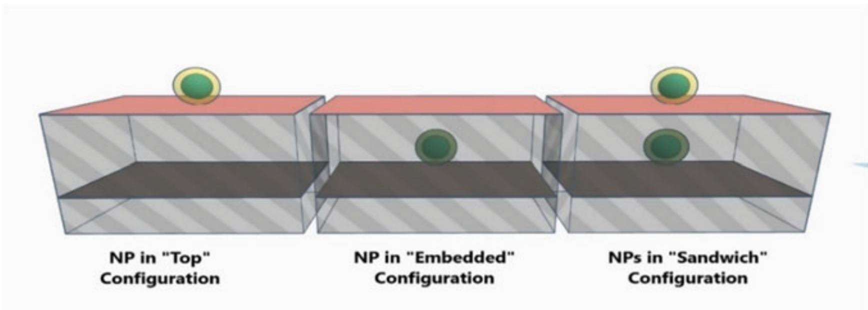


Fig. 1 Various configurations of a-Si substrate modified with Al-silica core–shell NP

Artificial Intelligence (AI), a transformational technology, is considered one of the foremost technologies driving the 4th Industrial Revolution through machine learning [8–11]. AI predicting behaviors of machines is expected to optimize machines for optimal performance, ensuring proper growth in businesses with optimal machine maintenance, reduced labor, better productivity, and balanced energy demand. The Genetic Algorithm (GA) utilized in this paper is an AI-based method for the optimization of any problem with wide nonlinear multidimensional searching spaces of potential solutions [8–11]. Jalali et al. developed a GA code by utilizing an appropriate fitness function to optimize some nanostructures that were used for multilayer antireflection coating on the surface of a solar cell substrate [10]. Later, Forestiere et al. utilized GA to optimize the design parameters of plasmonic nanoparticle arrays with the help of FDTD solver based on multiparticle Mie theory [11]. Additionally, they designed a surface-enhanced Raman Scattering (SERS) substrate by using the proposed GA optimization approach, where it demonstrated the importance of radiative coupling to strengthen the near-field due to the presence of nanoparticles [11]. Thus, Genetic Algorithm (GA) was chosen among other methods available to optimize a fitness function that implicates the aforementioned parameters. A fitness function for the GA code was developed with the aid of a predetermined coding scheme. The coding pattern was designed to consider the nanoparticle location and morphology, such as core diameter, shell thickness, and their material property.

2 Design and Procedure

The simulation setup (Fig. 2) and simulation environment were constructed and maintained to replicate real-life scenario, as explained elsewhere [12–16]. Procedure of the experiment involved aluminum-silica core–shell nanoparticles placed on top of the amorphous Si substrate, while varying the diameter of the Al core from 100 to 220 nm. Shell thickness (i.e., thickness of the silica coating) was varied from 2 to 20 nm. It should be noted that some larger particle sizes could not be investigated due to computational limits. The optimal nanoparticle of top configuration was determined using various performance parameters of TFSCs, explained below. The procedure was repeated later with the NPs embedded into the Si substrate at a depth of 2 nm from the top surface of the Si substrate. Next, the Al-silica core–shell NPs were positioned in a "sandwich" configuration. The size of the Al-silica core–shell NP on top was kept constant while varying the size of the NP that was embedded. The depth for "sandwich" configuration was also maintained at 2 nm.

To determine the optimum placement of NPs in a sandwich configuration, the depth of the embedded NP was later varied. The distance from the surface of the Si substrate to the top boundary of the shell of NP was varied from 2 to 20 nm for this purpose. This experiment demonstrated the coupling performance of electromagnetic radiation between the two nanoparticles in the "sandwich" configuration. To calculate optical and electrical parameters such as optical absorption enhancement factor (g), short-circuit current density (J_{sc}), and open-circuit voltage (V_{oc}),

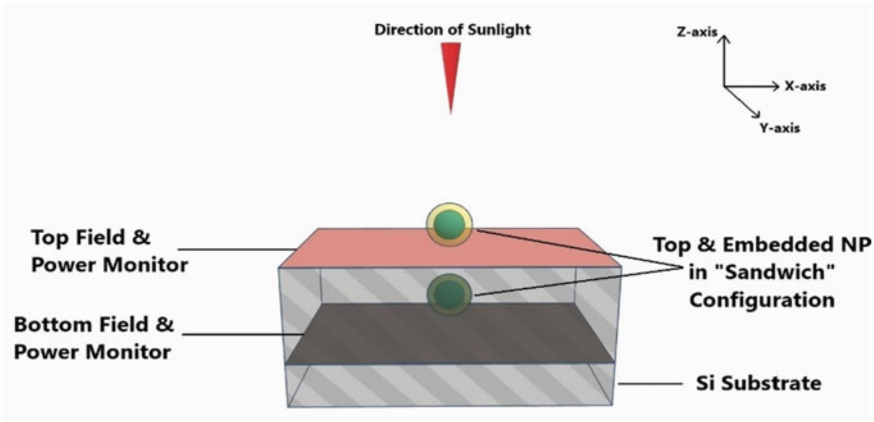


Fig. 2 Illustration of the FDTD simulation setup of "sandwich" configuration of Al-silica core–shell nanoparticles for optical and electrical analysis, using two top and bottom field and power monitors. Direction of incident light is set along the *z*-axis. Other variations of simulations included single Al-silica core–shell nanoparticle on top of the silicon substrate and single Al-silica core–shell nanoparticle embedded inside the silicon substrate

simulations were done for all size and configurations of nanoparticles. Using these values, other parameters such as fill factor (FF), power output (P_{out}), efficiency (η), and change in percentage efficiency ($\%\Delta\eta$) were calculated [14]. The optical absorption enhancement factor (g) was calculated using the equation given below which is defined as [14]

$$g = \int_{\lambda_1}^{\lambda_2} h(\lambda) d\lambda \tag{1}$$

where g is the sum of optical absorption enhancement factor across 150 frequency points within the wavelength $\lambda_1 = 250$ nm and $\lambda_2 = 750$ nm and $h(\lambda)$ is defined as

$$h(\lambda) = \frac{\text{absorption across Si substrate with plasmonic NP at a wavelength}}{\text{absorption across bare Si substrate at that wavelength}} \tag{2}$$

The short-circuit current density J_{sc} and V_{OC} values were calculated as shown previously [16, 17]. The fill factor (FF) is calculated using the result of the normalized open-circuit voltage V_{OC}(n) [17] found from electrically simulated open-circuit voltage from DEVICE (CHARGE) [13]. With previous values, the output power per unit area (P_{out}) was obtained [17]. Finally, the efficiency, η, and the change in percentage efficiency ($\%\Delta\eta$) were calculated using the result of total power output as follows [16]:

$$\eta = \frac{J_{sc} V_{oc} \mathrm{FF}}{1000\ \mathrm{A/m^2}} \times 100\% \tag{3}$$

$$\%\Delta\eta = \frac{\eta \text{ of substrate with NP} - \eta \text{ of bare substrate}}{\eta \text{ of bare substrate}} \times 100\% \tag{4}$$

Additionally, by positioning a frequency domain power and field monitor across the cross section of the core–shell nanoparticles and substrate, optical near-field images were generated to have a better visual understanding of the FDTD results.

The Genetic Algorithm (GA) optimizer model was developed in MATLAB. Firstly, initial populations are generated randomly with core at various diameters, silica shell thicknesses, and embedded nanoparticle depths (substrate) using appropriate mapping approach. Second, these populations were encoded using the binary encoding method (0 or 1). The first two bits on the left identify the location of the core; the following three bits indicate the depth of the core–shell; the next two bits indicate the shell thickness; and the final three bits indicate the core diameter. The binary encoding used to represent each chromosome in the population is depicted in Fig. 3, which is a flowchart for the GA algorithm developed is shown in for this application. Then, each member of the initial population is evaluated using a function referred to as the "fitness function". The fitness function includes core at various diameters, silica shell thicknesses, and NP depths (substrate) and utilized the fitness value as selection criteria for determining the "goodness" of each individual chromosome in the population. In this case, the widely used Roulette Wheel selection technique is utilized. Individuals are chosen based on their likelihood of being chosen. Using crossover and mutation process (rate = 0.5), offsprings are created which represents core at various diameters, silica shell thicknesses, and NP depths (substrate). The identical procedures will be performed on the new population, and the algorithm terminates when optimal configuration is identified. The mean square error is used to determine the difference between the expected and actual values of J_{sc}.

3 Result and Analysis

3.1 *Opto-Electronic Performance Analysis*

Extinction spectra of a single homogeneous Al nanoparticle and three Al-silica core–shell nanoparticles with different shell thickness are demonstrated in Fig. 4. The image depicts numerous peaks in the extinction spectra for all the nanoparticles shown. This is due to the metal nanoparticle being relatively bigger in size, leading to higher order multi-pole peaks in addition to the dipolar peak [18]. The thin-film solar cells modeled for the simulations in this study were traditionally used amorphous silicon (a-Si). The optical absorption range of the a-Si in this TFSC is represented by the gray region of Fig. 4, spanning from 250 to 750 nm. Figure 4 shows a homogeneous 100 nm aluminum particle (i.e., an aluminum nanoparticle with no silica

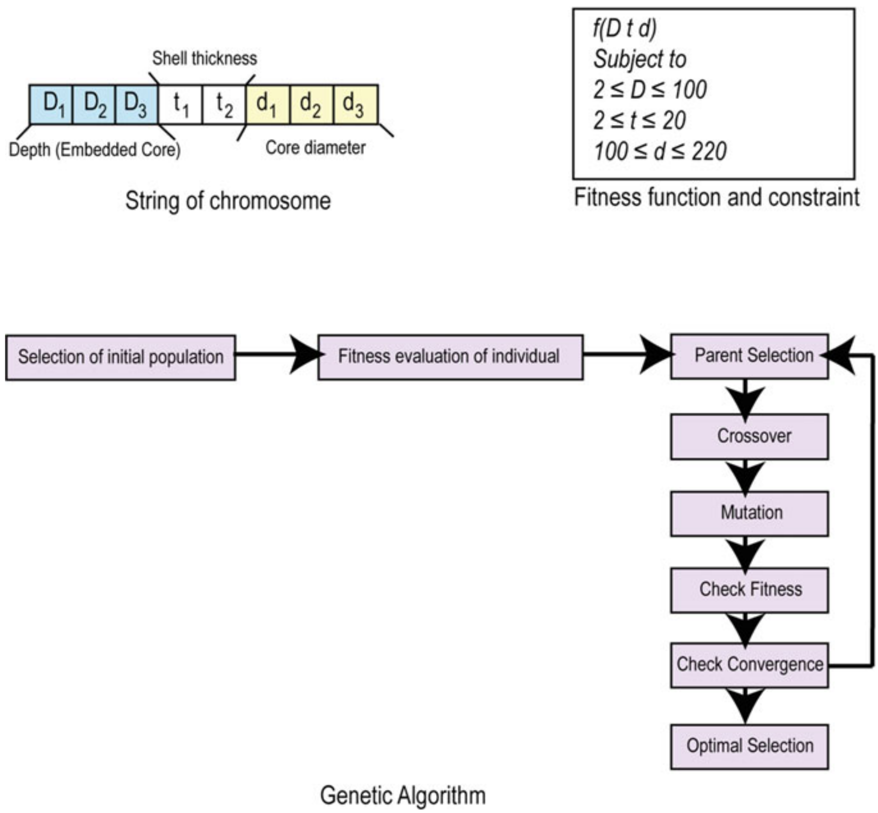


Fig. 3 Flowchart showing the Genetic Algorithm (GA) optimizer model used to predict and validate the results of the FDTD simulations

coating around it) with an extinction spectra peak wavelength of around 300 nm. The dotted blue curve, on the other hand, has a peak wavelength of 350 nm (i.e., an aluminum nanoparticle with a 10 nm silica coating around it). Since a substantial percentage of the absorbance spectrum of a-Si is more toward the longer wavelengths, the plasmonic resonant frequency of the Al NPs is not optimal for coupling with the amorphous silicon substrate (i.e., is more red-shifted). Therefore, to make the coupling between the silicon substrate of the thin-film solar cells and the Al nanoparticles more amenable, the metallic NPs are coated with a dielectric layer made of silica, as stated earlier.

Table 1 illustrates the computed opto-electronic performance parameter values for an absorbent silicon substrate with a single aluminum-silica core–shell NP positioned on top of the silicon layer, with various aluminum nanoparticle core diameters and silica shell thicknesses. The performance parameters are optical absorption enhancement (g), short current density (J_{sc}), open-circuit voltage (V_{oc}), maximum power output (P_{out}), efficiency, and percentage change in efficiency. A ‘bare’

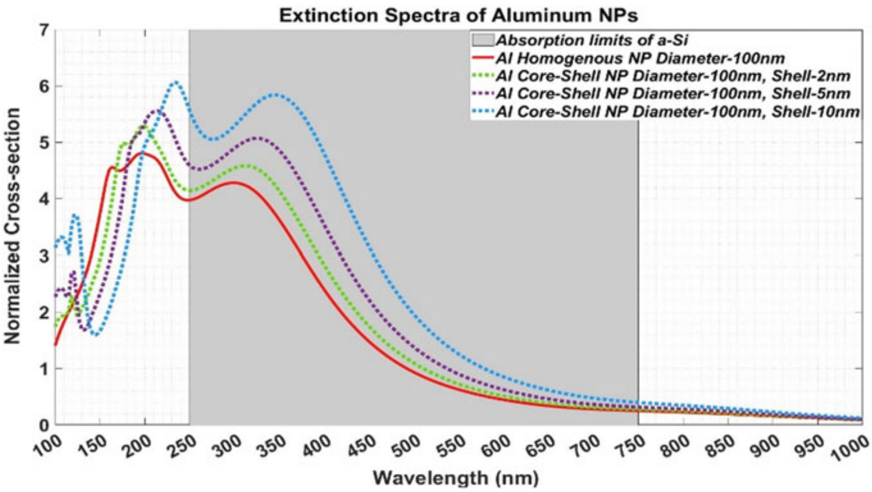


Fig. 4 Extinction spectra plots of homogeneous aluminum particle, core–shell aluminum-silica particles for varying shell thickness sizes (2, 5 and 10 nm), with absorption limits of a-Si

substrate refers to a silicon substrate without the presence of a nanoparticle. For each diameter (100, 120, 150, 200 and 220 nm), the shell thickness sizes (2, 5, 10, 20 nm) were varied. It is worth noting that the optical enhancement factor for a bare silicon substrate is 150. The simulation findings show that a shell thickness size of 2 nm generated the maximum performance in terms of g for each core size of aluminum nanoparticle, apart from 120 nm diameter NP, where a 5 nm shell size produced the best result of g. In addition, the nanoparticle with a core diameter of 150 nm and a shell thickness width of 2 nm had the greatest g factor of 280.37. If other parameters (e.g., recombination rates) are ignored, this NP was anticipated to generate the highest current from the thin-film solar cell as well, when compared to all the other models. The expected outcome is seen when the J_{sc} values were analyzed as given on Table 1. For the bare Si substrate, the J_{sc} value was 5.10 A/m^2. In comparison with the bare substrate, several of the Aluminum-silica core–shell nanoparticles placed on top of the TFSC obtained over a 100 percent rise in J_{sc}. The J_{sc} values obtained for core sizes of 150, 200, and 220 nm maintain the same trend as the g numbers as formerly narrated. Moreover, as previously expected, the NP with a core diameter of 150 nm and a shell thickness size of 2 nm has the highest J_{sc} value of 12.45 A/m^2. Nonetheless, not similar to the g numbers, the best J_{sc} findings were achieved for 100 nm core diameter and 120 nm core diameter with a shell thickness size of 20 nm rather than a narrower shell thickness. Various rates of electron–hole recombination in varying configurations of such plasmonic thin-film solar cells might be one cause for this.

The most significant improvement is portrayed by the percentage change in efficiency, which shows that 152.17% enhancement was produced for the 150 nm diameter and 2 nm shell Al-silica NP. Furthermore, for almost all smaller size NPs (i.e.,

Table 1 Opto-electronic performance parameters of Al-silica core–shell NPs of various diameter and shell thickness in "top" configuration

Diameter (nm)	Shell (nm)	g	J_{sc} (A/m^2)	V_{oc} (V)	P_{out} (W/m^2)	η	% Δ η
Bare Si		150	5.1	0.4	0.41	0.0414	0
100	2	236.14	10.4	0.408	0.86	0.0865	108.81
	5	218.38	9.9	0.407	0.82	0.0822	98.53
	10	224.60	10.3	0.404	0.85	0.0848	104.76
	20	232.63	10.6	0.404	0.88	0.0878	112.02
120	2	239.52	10.9	0.409	0.91	0.0907	119.1
	5	269.26	12.2	0.411	1.02	0.1019	145.99
	10	245.63	11.2	0.404	0.93	0.0927	123.79
	20	261.78	11.5	0.409	0.96	0.0961	132.01
150	2	280.37	12.5	0.411	1.04	0.1044	152.17
	5	277.26	12.2	0.411	1.02	0.1019	146.14
	10	244.85	10.7	0.404	0.88	0.0885	113.68
	20	244.95	10.1	0.408	0.84	0.0842	103.25
200	2	225.38	9.7	0.407	0.8	0.0804	94.09
	5	215.81	7.6	0.404	0.63	0.0630	52.11
	10	199.60	9.5	0.404	0.78	0.0784	89.36
	20	203.26	8.4	0.406	0.7	0.0696	68.07
220	2	224.05	9.5	0.407	0.79	0.0789	90.64
	5	200.89	8.8	0.404	0.73	0.0729	76.13
	10	190.15	8.9	0.404	0.74	0.0738	78.25
	20	160.67	7.4	0.404	0.61	0.0613	47.96

< 200 nm) that were simulated, the efficiency was greater than 100%, no matter the size of the shell thickness, as given in Table 1. This shows great promise for models of "plasmonic TFSCs" modified with Al-silica core–shell NPs placed on top of the silicon substrate.

The opto-electronic performance parameter values for a "plasmonic" thin-film solar cell with one aluminum-silica core–shell NP embedded into the silicon layer ("embedded" configuration), with different physical parameter, are given in Table 2. The performance metrics were kept similar to the simulation models given in Table 1, silica coatings of different shell thickness sizes (2, 5, 10, and 20 nm) were varied for each of the five diameters (i.e., 100, 120, 150, 200, and 220 nm). As stated earlier, absorption g factor for a bare substrate is 150 and J_{sc} of 5.06 A/m^2. When the shell thickness size is 20 nm, all the sizes of nanoparticles produced the highest g values along with nanoparticle with diameter 150 and 20 nm shell thickness generating the maximum g value. When it comes to J_{sc}, aluminum core with diameter of 220 nm and shell thickness of 2 nm produced highest J_{sc} of 7.81 A/m^2 for embedded configuration. In comparison, NPs implemented on top of the substrate produce a enhancement of 152.17% (Table 1) over NPs implemented inside the substrate with an enhancement of 55.33% (Table 2) (in contrast to the bare substrate). Thus, "embedded" configuration is considered non-significant when compared to "top" configuration in terms of plasmonic efficiency.

To further improve the performance of the "plasmonic" thin-film solar cells, core–shell Aluminum-silica nanoparticles were assembled in a "sandwich" configuration to improve on the findings achieved in Tables 1 and 2. Tables 3 illustrates the TFSC's g and J_{sc} numbers when the nanoparticles were arranged in a "sandwich" configuration (i.e., a single aluminum-silica NP on top of the silicon substrate with another single aluminum-silica nanoparticle embedded inside the substrate). Note that the g value was 150 and the J_{sc} value was 5.06 A/m2 for the bare substrate. The nanoparticle on top of the silicon substrate was maintained at a constant diameter and shell thickness size in the sandwich configuration, which was the most advantageous configuration derived from Table 1, where the NP core diameter was 150 nm and shell thickness of coating 2 nm. The core diameter of the embedded aluminum NPs was varied, where the optimum shell thickness size for each core diameter was utilized from

Table 2 Opto-electronic performance parameters of Al-silica core–shell NPs of various diameter and shell thickness in "embedded" configuration

Diameter (nm)	Shell (nm)	g	J_{sc} (A/m^2)	V_{oc} (V)	P_{out} (W/m^2)	η	% Δ η
Bare Si		150	5.1	0.4	0.41	0.0414	0
100	2	185.28	5.6	0.401	0.46	0.0463	11.84
	5	174.97	6.4	0.400	0.52	0.0524	26.58
	10	183.69	6.0	0.399	0.49	0.0491	18.60
	20	199.33	5.4	0.398	0.44	0.0441	6.42
120	2	192.34	5.1	0.399	0.42	0.0415	0.22
	5	183.05	6.4	0.400	0.52	0.0522	25.96
	10	195.29	5.8	0.399	0.48	0.0476	14.99
	20	237.19	4.0	0.397	0.33	0.0330	−20.40
150	2	208.94	6.1	0.402	0.50	0.0503	21.39
	5	194.54	6.3	0.400	0.52	0.0515	24.47
	10	211.55	5.6	0.399	0.45	0.0454	9.66
	20	261.00	3.8	0.397	0.31	0.0312	−24.68
200	2	233.16	6.7	0.397	0.55	0.0546	31.84
	5	214.15	6.2	0.400	0.51	0.0512	23.60
	10	230.76	5.2	0.399	0.43	0.0427	3.12
	20	241.51	4.5	0.398	0.36	0.0363	−12.32
220	2	219.12	7.8	0.403	0.64	0.0643	55.33
	5	222.05	6.3	0.400	0.52	0.0515	24.36
	10	238.55	5.1	0.398	0.42	0.0420	1.42
	20	244.35	4.3	0.398	0.35	0.0350	−15.54

Table 3 Optical absorption enhancement, g and short current density, J_{sc} of thin-film solar cells modified with nanoparticles in a "sandwich" configuration with varying core and shell sizes of embedded Al-silica NPs

Diameter (nm)	Shell (nm)	g	J_{sc}(A/m^2)
100	5	330.69	12.4
120	5	348.98	12.1
150	5	374.42	11.9
200	2	407.45	13.3
220	2	415.60	13.4

Table 4 Performance parameters of thin-film solar cells modified with NPs in a "sandwich" configuration with "top" NP diameter-150 nm, shell-2 nm and "embedded" NP diameter-220 nm, shell-2 nm with varying depth distances

Depth, nm	g	J_{sc} (A/m^2)	V_{oc} (V)	FF	P_{out}(W/m^2)	η	% Δ η
Bare	150	5.1	0.400	0.205	0.41	0.0414	0.00
2	415.60	13.4	0.412	0.204	1.13	0.1127	172.09
5	423.29	13.5	0.413	0.204	1.14	0.1136	174.29
10	422.85	13.2	0.412	0.204	1.11	0.1111	168.34
20	392.39	12.4	0.410	0.204	1.04	0.1038	150.74

Table 2. The lowest g value was obtained for 100 nm core diameter of embedded NP, whereas the lowest J_{sc} was produced for 150 nm core diameter of the embedded nanoparticle. For both these cases, the shell thickness size was 5 nm. The optimal result for both g and J_{sc} was produced when the core diameter was 220 nm, and the shell thickness size was 2 nm (where g was 415.60 and J_{sc} was 13.4 A/m^2) for the embedded NP in "sandwich" configuration.

Later, the depth distance of the "sandwich" arrangement, in which the embedded Aluminum-silica core–shell nanoparticles were inserted into the Si absorbing layer at different depths from the top surface of the silicon substrate, was investigated. This was accomplished by varying the spacing between the embedded nanoparticle and the surface of the substrate. Table 4 summarizes the findings.

3.2 *Optical Near-Field and Genetic Algorithm Analysis*

Optical near-field enhancement images demonstrate electromagnetic field distribution pattern in the near-field region of the plasmonic nanoparticle and the absorbing substrate. These images are generated by dividing the near-field data of the substrate with nanoparticle with the data of bare substrate. Thus, they show the interaction of the nanoparticle with the substrate when subjected to resonant electromagnetic wave. For this study, near-field enhancement image of the optimum sandwich configuration was generated at resonant wavelength = 440 nm, at which top 150 nm diameter NP demonstrated highest light coupling efficiency. Figure 5 portrays the optical near-field enhancement image of Al-silica core–shell nanoparticles in sandwich configuration with top NP diameter (D) = 150 nm, shell thickness (S) = 2 nm, embedded NP diameter (D) = 220 nm, and shell thickness (S) = 2 nm at Depth = 5 nm. The color bar represents linear scale, where black region demonstrates regions no enhancement and white (warm) region represents strong electromagnetic field distribution due to high light-matter coupling. As can be seen, sandwich configuration of Al-silica core–shell nanoparticle is resulting high electric field intensity in the near vicinity of the nanoparticle on top, represented by the white regions. This can explain the high electrical and optical performance of sandwich configuration when compared to other models.

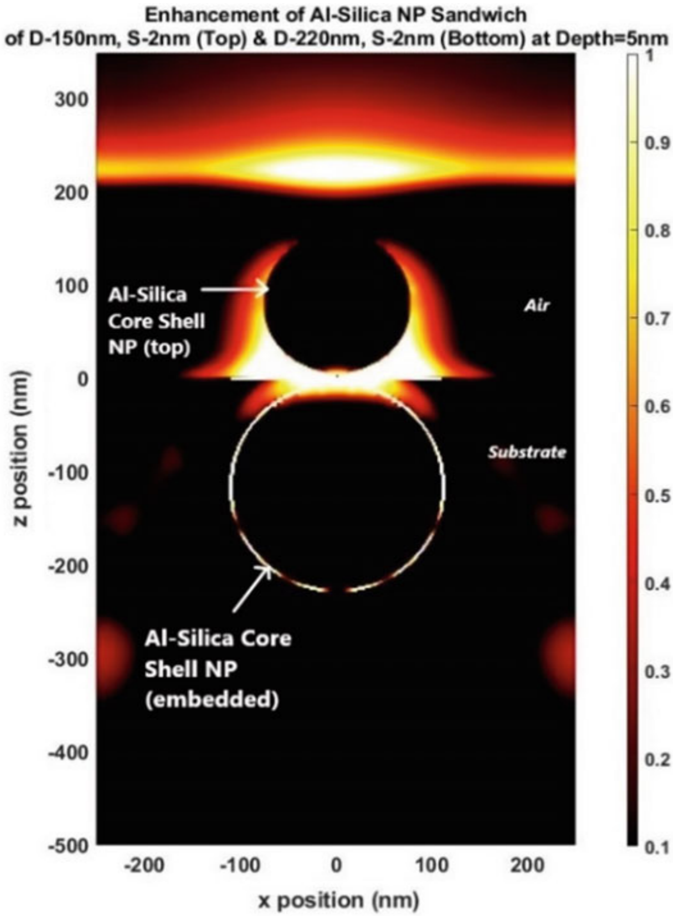


Fig. 5 Optical near-field enhancement image in x–z plane depicting EM field distribution pattern for Al-silica core–shell NPs in sandwich configuration

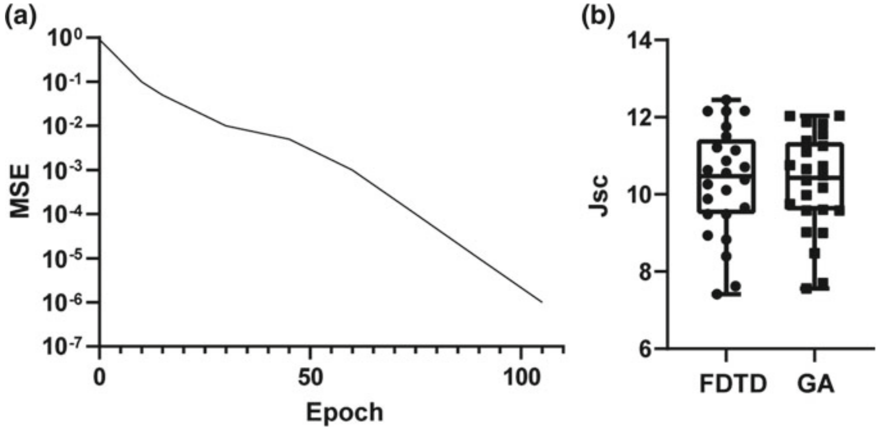


Fig. 6 **a** Plot of mean squared error (MSE) of results running Genetic Algorithm (GA) optimizer model as a function of iteration and **b** Plot of the spread of J_{sc} values generated by FDTD and predicted by GA optimizer model

Figure 6a shows the mean square error (MSE) as a function of iteration on running the Genetic Algorithm (GA) optimizer model. The iteration is run for the 110 epochs. It has been found that the MSE decreases with the iteration and reach to acceptable limit after 20 iterations. Figure 6b shows the box plot of J_{sc} for FDTD and GA. The mean square error is used to determine the difference between the expected and actual values of J_{sc}. The effect of core diameter and shell thickness on J_{sc} is shown in Fig. 7a and b for "top" and "embedded" configurations, respectively. It has been observed that the GA has selected optimal value of parameters such as position of shell, shell thickness, and code diameter. The data from FDTD calculation follows the same pattern as GA model result with small margin of error. This case is consistent with the study conducted by Jalai et al. [10].

4 Conclusion

In this study, an automated design procedure employing Genetic Algorithm (GA) integrated with finite-difference time-domain (FDTD) calculations has been used to show that different configurations of aluminum-silica core–shell nanoparticles can enhance the opto-electronic performance of thin-film solar cells. A "sandwich" configuration consisting of a first nanoparticle of diameter of 150 nm Al core with 2 nm thick silica shell layer placed on top of the absorbing Si substrate, and a second nanoparticle with a 220 nm Al core and 2 nm shell layer embedded inside the Si absorbing layer generated the most favorable TFSC performance parameter among the different configurations of nanoparticles studied. It has been postulated that interparticle coupling of Al-SiO_2 nanoparticles in a "sandwich" configuration

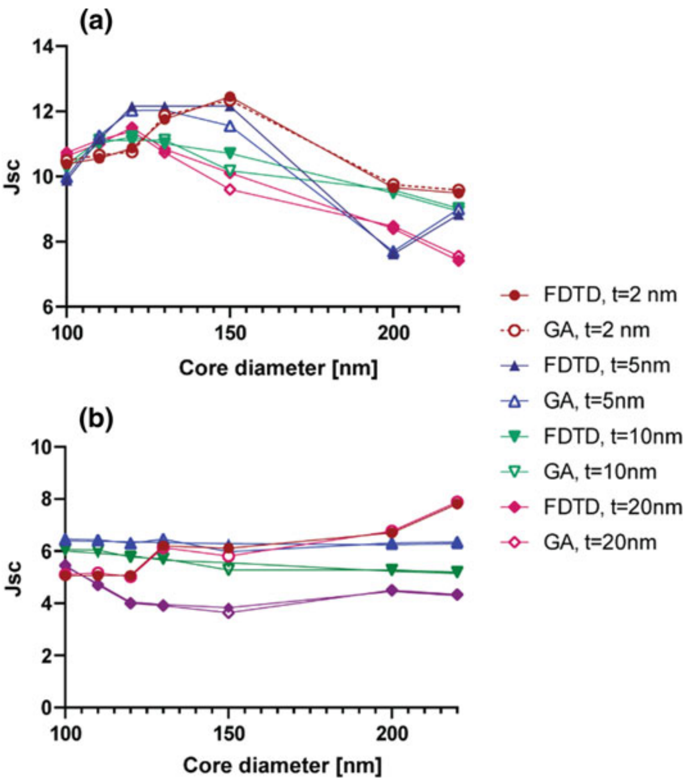


Fig. 7 Plot showing the effect of aluminum core diameter and silica shell thickness on J_{sc} values generated by FDTD and predicted by GA optimizer model for **a** "top" and **b** "embedded" configurations, respectively

result in the enhanced performance of TSFCs, compared to a bare Si substrate. The results of the FDTD calculations showed close agreement with predictions made of the performance of such "plasmonic" TFSCs made by a GA optimizer model. This is consistent with observations from the previous studies where GA results follow the same trend as experimental data with low relative error. These promising observations can be further explored to open the door for future collaborative use of GA and FDTD calculations in a comprehensive manner to design high sensitivity photovoltaic devices.

Acknowledgements The authors would like to thank Independent University, Bangladesh (IUB), for funding this research (Research Project No: 2020-SETS-04) and Jahangirnagar University for other logistical support.

References

1. Breeze, P.: Power generation technologies. Newnes (2019)
2. Catchpole, K.A., Polman, A.: Plasmonic solar cells. Opt. Exp. **16**(26), 21793–21800 (2008)
3. Dreizin, E.L.: Metal-based reactive nanomaterials. Prog. Energy Combust. Sci. **35**(2), 141–167 (2009)
4. Maier, S.A.:Plasmonics-fundamentals and applications. Springer (2007)
5. Schatz, G.C., Van Duyne, R.P.: Electromagnetic mechanism of surface-enhanced spectroscopy. Handb. Vib. Spectrosc. (2006). https://doi.org/10.1002/0470027320.s0601
6. Jensen, T., Kelly, L., Lazarides, A., Schatz, G.C.: Electrodynamics of noble metal nanoparticles and nanoparticle clusters. J. Cluster Sci. **10**(2), 295–317 (1999)
7. Movsesyan, A., Baudrion, A., Adam, P.: Extinction measurements of metallic nanoparticles arrays as a way to explore the single nanoparticle plasmon resonances. Opt. Exp. **26**(5), 6439 (2018). https://doi.org/10.1364/oe.26.006439
8. Alì, G., Butera, F., Rotundo, N.: Geometrical and physical optimization of a photovoltaic cell by means of a genetic algorithm. J. Comput. Electron. **13**(1), 323–328 (2013). https://doi.org/10.1007/s10825-013-0533-0
9. Shapiro, J.: Genetic algorithms in machine learning. In: Paliouras, G., Karkaletsis, V., Spyropoulos, C.D. (eds.) Machine Learning and Its Applications. ACAI 1999. Lecture Notes in Computer Science, vol. 2049. Springer, Berlin, Heidelberg (2001). https://doi.org/10.1007/3-540-44673-7_7
10. Jalali T.J., Marjan Mohammadi, A.: Genetic algorithm optimization of antireflection coating consisting of nanostructured thin films to enhance silicon solar cell efficacy. Mater. Sci. Eng. B. **247**, (2019). https://doi.org/10.1016/j.mseb.2019.05.016
11. Forestiere, C., Pasquale, A.J., Capretti, A., et al.: Genetically engineered plasmonic Nanoarrays. Nano Lett. **12**, 2037–2044 (2012). https://doi.org/10.1021/nl300140g
12. Ansys-Lumerical: www.lumerical.com/tcadproducts/fdtd/
13. Ansys-Lumerical: www.lumerical.com/products/#device
14. Shaky, M., Chowdhury, M.: Bowtie-based plasmonic metal nanoparticle complexes to enhance theopto-electronic performance of thin-film solar cells. Appl. Opt. **60**, 5094–5103 (2021)
15. Gupta, A., Kumar, P., Pachauri, R., Chauhan, Y.K.: Effect of environmental conditions on single and double diode PV system: comparative study. Int. J. Renew. Energy Res. (IJRER) **4**(4), 849–858 (2014)
16. Arefin, S.M.N., Islam, S., Fairooz, F., Hasan, J., Chowdhury, M.H.: Computational study of "sandwich" configuration of plasmonic nanoparticles to enhance the optoelectronic performance of thin-film solar cells. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1237–1240 (2020). https://doi.org/10.1109/TENSYMP50017.2020.9230633
17. Fairooz, F., Hasan, J., Arefin, S.M.N., Islam, S., Chowdhury, M.H.: Use of plasmonic nanoparticles made of aluminum and alloys of aluminum to enhance the optoelectronic performance of thin-film solar cells. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1233–1236 (2020). https://doi.org/10.1109/TENSYMP50017.2020.9230672
18. Bohren, C.F., Huffman, D.R.: Absorption and Scattering of Light by Small Particles. Wiley-VCH, Berlin, Germany (2004)

Explainable Sentiment Analysis for Textile Personalized Marketing

Mukwaya Kasimu, Nakayiza Hellen, and Ggaliwango Marvin

Abstract According to the United Nations (UN), the 2030 Agenda recognizes ending poverty in all of its forms and dimensions, including severe poverty which is the biggest global challenge and a necessity for sustainable development. Sustainable Development Goal (SDG) 1 aims to "end poverty in all its forms everywhere" but a global uniform approach has not been found rather than country-specific programs that tackle poverty with international initiatives to support them toward the SDG poverty eradication goal. With the opportunities generated by the 4th Industrial Revolution, nations can optimize their economic strengths to end poverty. Since the textile industry is one of the economic power engines of many developing countries, optimal strategic enhancements can feasibly ignite unexplored potentials with the emerging 4th Industrial Revolution technologies. With the rapid technological infrastructural developments, cloud computing has powered up lots of e-commerce businesses that enhance distribution of textile products. The problem is that there is no feedback channel for local and global customers to share their experience with the textile product developers and distributors, and this limits the marketing potentials of the textile producers and distributors locally and globally. In this work, we propose and demonstrate an explainable data-driven solution that optimizes 4th industrial Artificial Intelligence potentials and Natural Language Processing to boost textile product development and distribution. Our proposed Artificial Intelligence approach to sentiment analysis transparently explains evolving customer sentiments and preferences for textile products and provides insights for textile product development, personalized marketing, and distribution of textile garments for women.

Keywords Explainable Artificial Intelligence (XAI) · Sentiment analysis · Personalized marketing · Textile industry · Technology adoption and management

M. Kasimu (✉) · N. Hellen · G. Marvin
Brac University, Dhaka 1212, Bangladesh
e-mail: mukwaya.kasimu@g.bracu.ac.bd

N. Hellen
e-mail: nakayiza.hellen@g.bracu.ac.bd

G. Marvin
e-mail: ggaliwango.marvin@g.bracu.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_33

1 Introduction

The Textile and Clothing (T&C) sector is one of the world's oldest, largest, and most international industries. It is a labor-intensive "starter" business for countries engaged in export-oriented industrialization. T&C provides a variety of options, including entry-level positions in developing countries for unskilled labor. The technological characteristics of the T&C industry have made it suitable as the first step on the "industrialization ladder" in poor countries such as Bangladesh, Sri Lanka, Vietnam, and Mauritius, which have since become middle-income countries after experiencing rapid output growth in the sector (Vietnam, Mauritius).

There are a variety of reasons why the apparel industry has been so vital to economic progress. The sector consumes a huge amount of unskilled labor, which is typically drawn from rural farming households. Despite cheap start-up costs, the sector's expansion provides a foundation on which to generate capital for more technologically demanding activity in other industries. As the industry grows, more modern technologies can be imported and paid for with cash generated by garment exports.

The textile and garment industry is the preferred entry point for most emerging countries seeking to enter the industrialized world. The ease of entrance into this area, as well as the abnormally high pay in industrialized countries, has produced advantageous conditions for the production and exportation of textile and garment derived items. At the same time, this unique circumstance has fueled severe protectionism in many industrialized countries where export markets are present, resulting in a harsh struggle among the many participants. Surprisingly, it is the United States' trade policies that have aided the growth of the textile and garment industries in many countries.

Although the fast adoption of e-commerce, particularly business-to-business (B2B) e-commerce among the ready-made garments (RMG) industry is a crucial potential for developing nations like Bangladesh to develop their local and international commerce. Despite the great expectations for reaping the benefits of B2B e-commerce in RMG sectors, its adoption in these countries remains poorly understood and under-researched. Most of these countries, on the other hand, have abandoned meaningful examination of direct customer feedback, their experience with exported products, and all relevant data. The countries have rested it all on a B2B model, which has created a significant chasm between them and their product users, whose sentiments are mixed. More so for those who have employed the method have limited knowledge about AI and how to appropriately use it to join the rift between them and their consumers.

Consumer behavior and company practices have both changed as a result of the Internet and social media's usage. Organizations can benefit from social and digital marketing by lowering expenses, increasing brand awareness, and increasing revenues. Negative electronic word-of-mouth, as well as intrusive and annoying online brand presence, poses substantial issues and hence the need for sentimental analysis to fully explore customers' satisfaction, feelings, and their general experience. Sentiment analysis, call it opinion mining, is the strategy used to extract and evaluate sentiments expressed in textual data.

The rise of reviews, forum discussions, blogs, microblogs, Twitter, and social networks has increased the importance of sentiment analysis. For the first time in human history, we now have access to a massive amount of opinionated data in digital form that can be analyzed [1].

The 4th Industrial Revolution has expanded lots of technological infrastructures that are underutilized by other industries which do not adopt quickly. Some of them include cloud infrastructural technologies, Artificial Intelligence, and Internet of Things among others. The textile industry can potentially adapt to the rapid technological trends [2, 3]. With explainable sentiment analysis, we can understandably extract customer sentiments from textile reviews and ratings, textile social media data, and textile e-commerce platforms among others using abstracted machine learning algorithms. We can transparently comprehend the evolving customer sentiments and preferences to textile products which would iterate with changes in style and fashion trends. We can interpretably derive insights for strategic textile product development and improve textile customer service delivery besides textile distribution channels and services. Explainable sentiment analysis would also greatly help in solidifying the strategies for personalized marketing of textile products and this is a long-term strategy for conquering local and global textile markets [4–6]. Our contributions are

- We propose and demonstrate an explainable and understandable technological process of sentiment analysis that harmonizes with the 4th industrial Artificial Intelligence revolution to boost economic growth of nations' stakeholders in the textile industry.
- We propose and demonstrate use of machine learning to predict textile product recommendations based on customer reviews and ratings.
- We demonstrate an interpretable approach to derive insights to textile personalized marketing, textile strategic product development, and distribution of textile products with improved customer service.

The rest of the paper has been organized as—Existing Works in Sect. 2, Proposed Approach in Sect. 3, Results and Discussion in Sect. 4, and eventually Conclusion in Sect. 5.

2 Existing Works

The Internet has become ingrained in people's daily lives, allowing many online service companies to flourish. However, in order to continue their operations and remain competitive, these businesses must be concerned about the quality of the services they give. In this situation, it is critical to be able to gauge client satisfaction with such services and sentiment analysis is the support tool that can help to enrich customer satisfaction evaluations [7].

Billions of people throughout the world, the online world, social networking sites, smartphone applications, and other digital communications technology have become part of their daily lives. According to January 2020 data, 4.54 billion people utilize the Internet, accounting for 59% of the worldwide population. In 2019, there were 2.95 billion active social media users globally. By 2023, this number is expected to reach over 3.43 billion. Digital and social media marketing may help businesses achieve their marketing objectives at a low cost. More than 50 million companies have Facebook pages, and 88% of enterprises use Twitter for marketing. These locations are frequently used to increase public awareness of government services and political messages or rallies.

People are spending a growing amount of time online looking for information, discussing products and services with other customers, and interacting with businesses. Companies have reacted to these customer behavior changes by incorporating a crucial component in their advertising strategies which is to go digital.

Organizations like textile firms must deliberately employ social media advertising to match with their intended audiences, with increased customers using social media on a regular basis for activities such as finding news, researching on products, and amusement [8]. The lack of appropriate scales to measure and investigate constructs of interest, the incessant t changes in current and emerging social network platforms, and its application analysis to research the flow of electronic speech, messages, and the impact of such information on consumers' point of view and behaviors is all difficulties in employing social media to satisfy customers [7]. Digital and Social Media Technologies (D&SMT) have become a part of lives of billions of people in the past 20 years; since then, D&SMT (e.g., e-commerce, online brand communities, digital advertising tactics, live chat services, mobile services) have transformed the engineering of captivating customer experiences by availing new ways to reach, inform, sell to, learn about, and serve customers with an overarching social dimension [9].

Understanding emotions is among the most crucial components of personal development and growth and is key to human intelligence emulation. The ability to gather public attitudes about social events, marketing campaigns, political movements, and product choices automatically has aroused public attention causing emergence of affective computing and sentiment analysis fields [9].

Smart textiles have been incorporated more in service ecosystems that go past the current horizontal textile value chain. This brings more opportunities for textile developers, product and service designers to develop close-to-the-body applications in the well-being area. The role taken on by the body, the level of personalization, and the prototyping process provides opportunities for ultra-personalization within these latest embodied STS varieties.

This will increase smart textiles' material and tangible features to include intangible properties from services, like the capability to quantify, store data while changing a material's functioning over time. As a result, it is becoming highly vital for textile developers and service providers to work more collaboratively so as to build smart textile service (STS) types [10].

Textile firms can offer personalized items at a minimal extra expense because of mass customization, which has inspired a lot of research on the supply side but little on customer views. Customers' willingness to pay more for mass-customized products may be influenced by their desire for distinctive items and the usage of such items in their personal branding. It could also represent their desire to avoid the drawbacks of standardized, off-the shelf items. These positive and negative impulses are not always correlated, implying a segmentation strategy. The proposed model incorporates additional antecedents derived from existing research in addition to these primary impacts. A wide sample of real-world consumers confirms the key predictions in testing of the conceptual model, revealing the existence of four client categories, each with a particular attitude toward customized products [11]. This, however, can only be achieved through analyzing sentiments of different segments.

Throughout the literature, there has been consistent evidence that many textile firms have gone digital and while there has been much research on how to fully utilize the Internet for profit maximization, there are challenges that come with the digital platforms, for example, limited knowledge.

In the technological context, the 4th Industrial Revolution comes with its own challenges that render most of the existing approaches obsolete. Thus, all companies that aim at globalizing their markets must adapt to the new technological trends and those that lag behind are all going to be knocked out of the market. Moreover, the new technological trend is founded on technologies that are not understandable by most of the stakeholders and participants in the global textile markets. In this paper, we aim at orienting the textile stakeholder and business investors to some of the technological trends that disrupts global markets and these include, Natural Language processing and Artificial Intelligence for electronic commerce. We also introduce the concept of data-driven textile personalized marketing based on customer experiences. Artificial Intelligence has to be explainable for easy adoption in such a way that economic power engines of developing countries can transparently and reliably guide on textile product development and distribution. This is the future of the textile industry in the 4th Industrial Revolutionized global markets.

3 Proposed Approach

We proposed an explainable and interpretable approach to textile sentiment analysis of reviews and ratings of textile industry stakeholders to understand the changes of their customer sentiments and comprehensively interpret the predictions of Machine Learning (ML) and Artificial Intelligence (AI) models for reliable strategic business insights. This approach synchronizes with AI emerging technologies of the 4th Industrial Revolution to enable textile business owners to understand the changes in sentiment and preferences of their garment customers.

3.1 Explain It like I'm 5(ELI5) Model Selection and Description

We used ELI5 Python package XAI Libraries to explain the predictions of ML to its built-in support for several ML frameworks [12]. ELI5 used weights associated with every feature to depict the feature's contribution to the final prediction power in each ML model for decreased ambiguity and better interpretability of the labels [13]. Since ELI5 shuffles the removed variable attribute values and randomizes the chosen variable to analyze the decrease in the model's performance, we used it to compute and interpret the global explainability of the selected models in analysis of the variable comment in the review text [14].

3.2 Dataset Feature and Description

In this study, we used women's clothing e-commerce reviews [15] as the dataset. This dataset comprises reviews given by anonymous real customers that is to say; customers' names were not included due to the fact that it involves real commercial data of the customers, and recommendations to the company were swapped with "retailer". Table 1 illustrates the frequency distribution of dataset features (with their descriptions) and the label (recommended IND).

The dataset had over 23,486 rows and 10 feature variables. Each row corresponded to a customer review and included the relevant variables.

Table 1 Frequency distribution of dataset features [15]

Feature	Description	Unique count
Clothing ID	Integer categorical variable that refers to the specific piece being reviewed	1172
Age	Positive integer variable of the reviewers age	77
Title	String variable for the title of the review	13,984
Review text	String variable for the review body	22,621
Rating	Positive ordinal integer variable for the product score granted by the customer from 1 worst to 5 best	5
Recommended IND	Binary variable stating where the customer recommends the product where 1 is recommended and 0 is not recommended	2
Positive feedback count	Positive integer documenting the number of other customers who found this review positive	82
Division name	Categorical name of the product high-level division	3
Department name	Categorical name of the product department name	6
Class name	Categorical name of the product class name	20

3.3 Data Preparation and Preprocessing

Text Cleaning: We began with cleaning the user review texts to remove delimiters found in the review texts for example \n and \r.

Sentiment Analysis: We used the NLTK [16] sentiment analyzer to automate the process by preparing text data for mathematical analysis instead tagging the user comments (review texts) manually. This helped in elimination of all intuitive tags of review texts with a 3 rating threshold such that positive feedback is based on a review rating greater than or equal to 3 and negative feedback otherwise. We used NLTK because the manual approach, intuitive tagging was insensitive to neutral sentiments [16] (Fig. 1).

A positive sentiment was detected among most reviews similar to the rating distribution. However, the occurrence of neutral rating was lower compared to medium-ranged occurrence ratings when it came to the distribution of the ratings. The plot on the bottom right of Fig. 2 shows that as the ratings got higher, positive sentiment reviews exhibited an increasing occurrence. Surprisingly, a rating of 3 was attained by neutral and negative and neutral sentiment reviews. Since 3 was the highest occurrence rating, it indicated that the customers' were often motivated to assign a review score.

In Fig. 3, the first top row demonstrates non-recommended reviews and lower row illustrates recommended reviews although the same variables were used. This helped us demystify the nature of recommended reviews depending on the mood of the writing together with customer assigned ratings. Surprisingly, the department distribution was not changing with status of recommendation but the rating was completely changed. Figure 4, illustrates how the variables influence each other. There is no linear relationship between the variables, therefore there is no multicollinearity. The maximum correlation is 79% between Rating and Recommended IND.

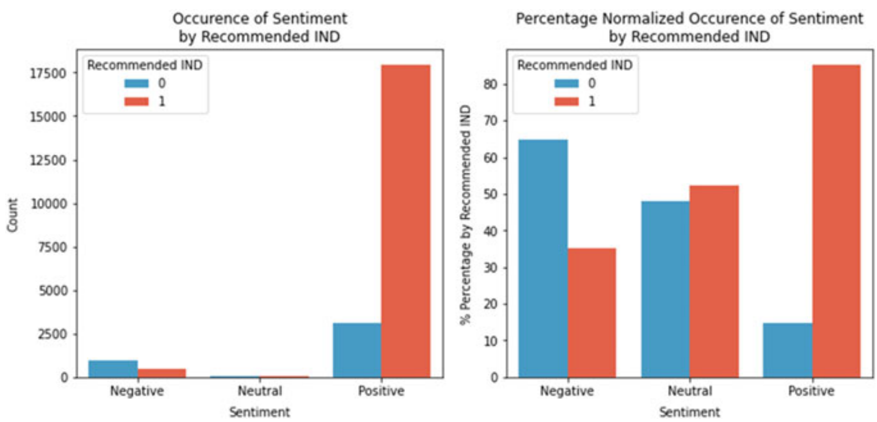


Fig. 1 Sentiment occurrence by recommendation indicator

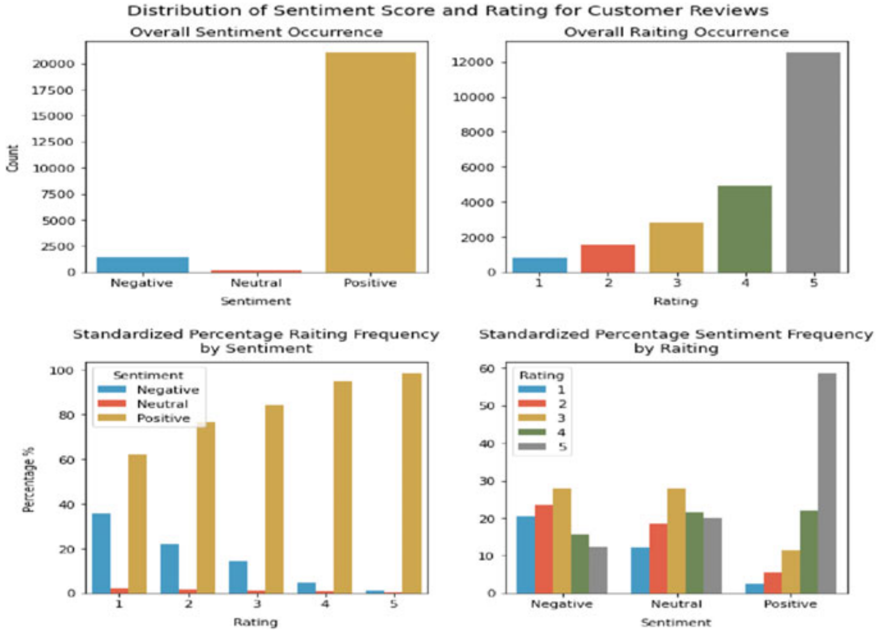


Fig. 2 Percentage rating by sentiment frequency and rating

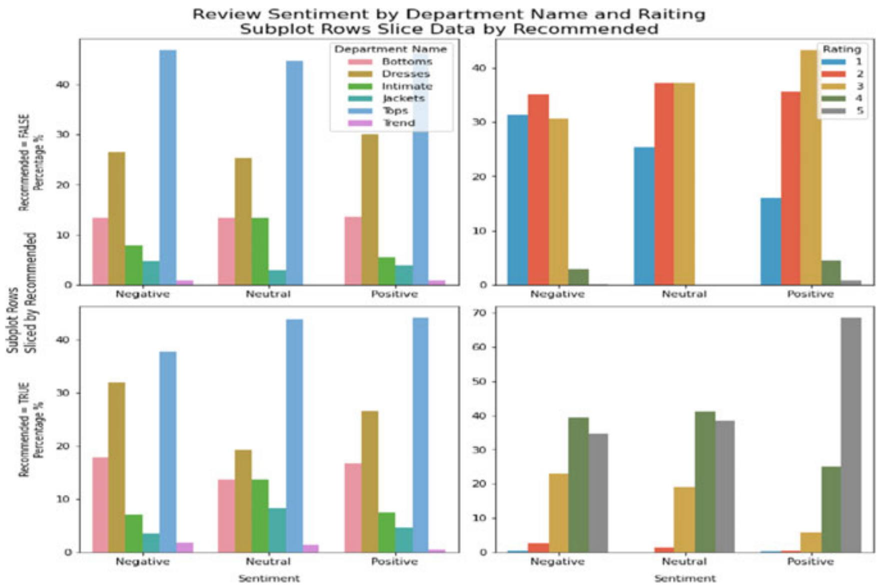


Fig.3 Review sentiment by department name and rating

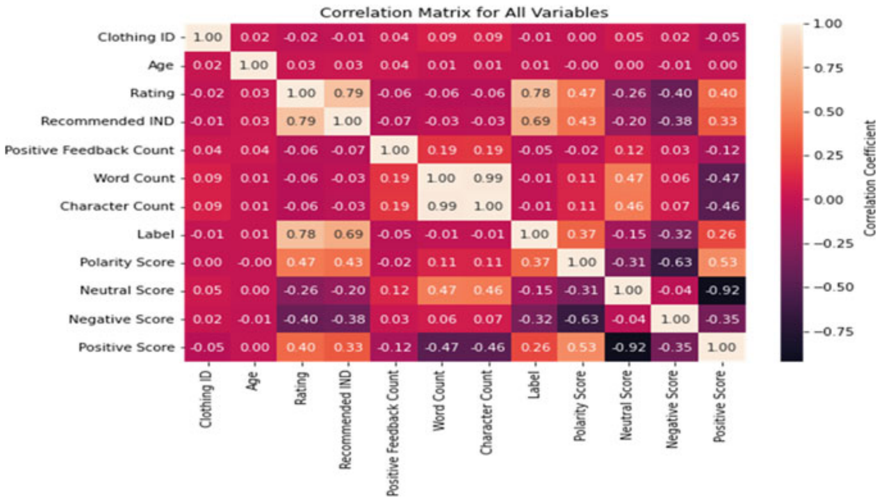


Fig. 4 Correlation matrix for all variables

Fig. 5 Correlation between positive feedback count and positive score

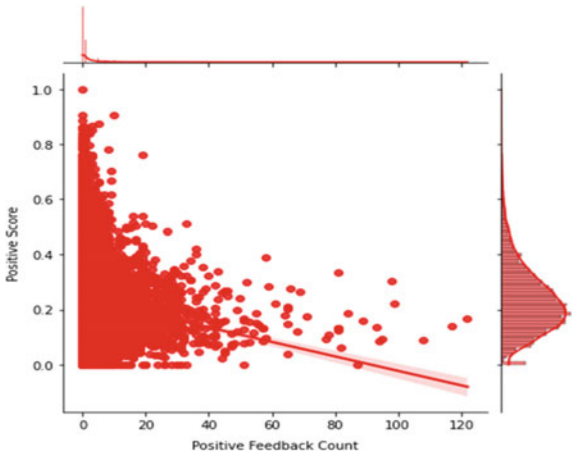


Figure 5 showed that there is a significant negative correlation between positive feedback count and positive score. It suggested that most of the customers probably gave just constructive criticism but not outright positivity.

3.4 Word Distribution and Word Cloud

Word cloud provides a graphic structure for the frequency of occurrence for specific words by text formatting and word frequency resolution. The graphic structure we

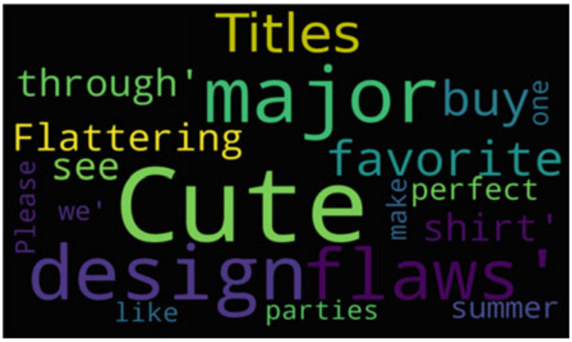


Fig. 6 Most frequent words for review titles

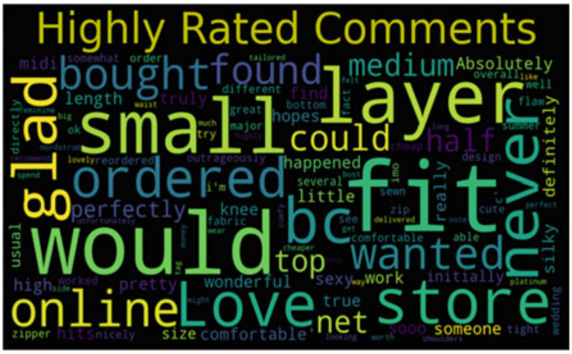


Fig. 7 Most frequent words in reviews

formed was filled with words such that larger sized text signified more frequency of word occurrence as shown in (Figs. 6 and 7).

Unfortunately, word clouds only demonstrated the distribution of individual text, and this removes the word context and disregards negative prefixes as well. To resolve this problem, we applied *n*-grams to increase the size of detected values from one word to several other words. This enabled demonstration of frequency counts among word arrangements (Fig. 9).

3.5 N-Grams by Recommended Feature

We realized that textile product inconsistency and cloth fit appeared as main subjects in the reviews. The core subjects in the textiles product reviews that were recognized by *n*-grams were: Fit (either the product's marketed size truly matched customer size and height), Hate or Love (customer feelings about the product), Complements

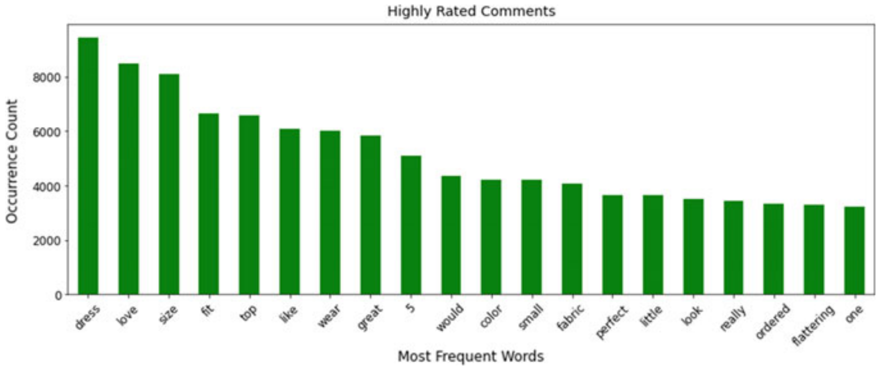


Fig. 8 Most frequent words in highly rated comments

Fig. 9 **a** Word clouds by department name. **b** Word clouds by department name

(customer experiences with the clothing), Product consistency (whether the clothes matches the marketed quality expectations) as visualized in (Figs. 10 and 11).

Non-Recommended Items

	1-Gram	Occurrence	2-Gram	Occurrence	3-Gram	Occurrence	4-Gram	Occurrence	5-Gram	Occurrence
0	dress	1976	wanted love	243	really wanted love	70	really wanted love dress	15	reference 5 7 125 lb	3
1	like	1780	going back	215	wanted love dress	65	looked like maternity top	10	going back wanted love dress	3
2	top	1572	looked like	187	really wanted like	40	really wanted like dress	9	reference measurements 38 30 40	3
3	would	1348	looks like	153	made look like	29	really wanted like top	9	photos reference measurements 38 30	3
4	fit	1327	really wanted	151	wanted love top	28	5 4 120 lbs	8	medium photos reference measurements 38	3
5	fabric	1245	made look	135	fit true size	28	really wanted love top	7	wearing medium photos reference measurements	3
6	size	1209	look like	126	ordered usual size	25	5 7 140 lbs	6	back really wanted like dress	3
7	back	1039	5 4	113	way much fabric	25	measurements 38 30 40	6	arm holes also cut low	3
8	look	965	felt like	109	sadly going back	24	going back really wanted	6	wanted top work looks cute	2
9	love	928	true size	104	like maternity top	23	would look better someone	5	sides much shorter front back	2
10	small	910	usually wear	103	usually wear size	23	ordered usual size small	5	115 lbs ordered xs petite	2
11	5	871	love dress	95	one going back	20	wanted love dress much	5	would expect retailer shirt also	2
12	really	869	much fabric	89	first time wore	18	really wanted love shirt	5	perhaps would look better someone	2
13	ordered	825	5 5	88	would look good	16	looks like maternity top	5	reference 5 2 24 waist	2
14	wear	756	fit well	81	looked like wearing	16	back really wanted like	5	petite reference 5 2 24	2
15	material	736	size small	80	looks great model	15	5 5 122 lbs	4	5 4 really wanted like	2
16	much	721	thought would	80	looked like maternity	15	made look like pregnant	4	returning item wanted love dress	2

Fig. 10 Gram occurrence of non-recommended items (partial snipped image)

Recommended Items

	1-Gram	Occurrence	2-Gram	Occurrence	3-Gram	Occurrence	4-Gram	Occurrence	5-Gram	Occurrence
0	dress	8591	true size	1243	fits true size	264	compliments every time wear	46	34b 26 waist 36 hips	27
1	love	8017	love dress	657	fit true size	192	26 waist 36 hips	32	get compliments every time wear	21
2	size	7561	5 4	622	received many compliments	163	34b 26 waist 36	28	5 2 currently 33 25	17
3	fit	5995	usually wear	588	runs true size	143	looks great skinny jeans	25	115 lbs 30 dd 26	15
4	top	5846	looks great	574	love love love	138	get compliments every time	23	2 currently 33 25 37	13
5	wear	5678	fit perfectly	553	usually wear size	107	love love love dress	22	115 lbs 30dd 26 5	13
6	great	5584	well made	531	every time wear	81	115 lbs 30 dd	22	lbs 30 dd 26 5	13
7	like	5368	love top	524	ordered usual size	79	usually wear size 4	21	30 dd 26 5 waist	12
8	5	4678	5 3	488	small fit perfectly	72	5 4 120 lbs	20	5 1 5 115 lbs	12

Fig. 11 Gram occurrence of recommended items (partial snipped image)

These reviews are often unhelpful because they spoil textile distributor's reputation yet the views on public online platforms are the most important assets. Otherwise, positive reviews have no criticism, they carry affirmative confirmation for fit therefore sharing this kind of experience with the textile product ("True Size", "Fit Perfectly", "Fit like a glove") besides multiple 2-g with customer's height proposes positive reviews to approve product fit depending on size. The frequent appearance positive reviews suggested that height and size were the main subjects' necessary retail customer consistency and satisfaction. "Received many compliments", "Look forward to wearing", "Every time I wear", "Looks great with jeans" are the reviews that reflected customer's experience with the clothing in public. This not only expressed the relevance of trends but also suggested that textile product reviews were a highly dependent on social space were customers talk about retailers and fellow customers too.

3.6 Intelligent Supervised Learning

Supervised learning needs features (independent variables) and labels (dependent variables). A tuple with the comment and customer rating label is created, and in this case, the entire comment acts as the independent variable. But for the algorithm to work, every word must be treated as a variable. Instead of applying sequential words, the model records words that are present in the entire word dictionary existing in the corpus of comments. For computational intensity reduction, only the top 5000 most common words were considered. Word presence for a piece of text was checked against word features in the dataset. This was done to every customer review using a loop while retaining every label of the reviewer.

Logistic Regression

This preferred simple classification model for multiple explanatory variables is given by

$$p\big(Y_i|X_i,\ldots,X_p\big)=\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_pX_p}}$$

where $X = (X1, \ldots, X_p)$ is p explanatory variables for predicting the response variable Y, and it can also be expressed as;

$$\log\left(\frac{p(Y_i|X_i, \ldots, X_p)}{1 - p(Y_i|X_i, \ldots, X_p)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

where the logit function of p exhibits linearity in explanatory variables for estimated multiple logistic regression models to classically predict the probability of a given observation as positive or negative for correct classification of new observations in untrained labels. Here, test error rate is reduced by assigning each observation to its most likely class trained on the values of the explanatory variables; hence, a test observation with explanatory vectors $x1, \ldots, xp$ should be assigned to the class j for which $p(Y = j|X_1 = x_1, \ldots, X_p = x_p)$ is largest. This corresponds to assigning a test observation to class 1 if $p(Y = 1|X_1 = x_1 \ldots, X_p = x_p) > 0.5$, and to class -1 otherwise in binary setting [2].

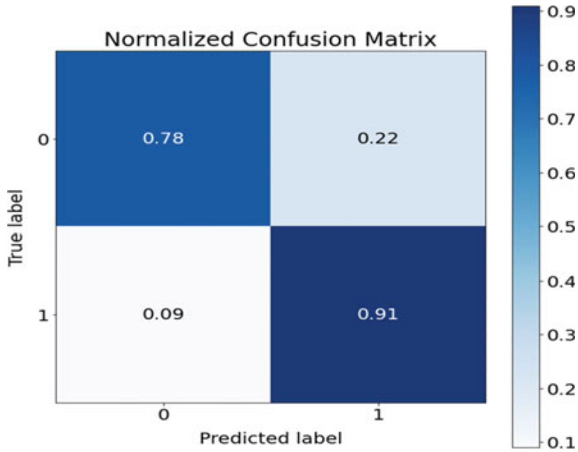


Fig. 12 Confusion matrix

Table 2 Classification report

Metric	Precision	Recall	F1-score	Support
0	0.56	0.78	0.65	596
1	0.96	0.91	0.94	3830
Accuracy			0.89	4526
Macro average	0.76	0.84	0.79	4526
Weighted average	0.91	0.89	0.90	4526
Accuracy score: 0.89				

Table 3 Training and validation results

Metric	Value
Train set accuracy	0.91
Train set ROC	0.89
Validation set accuracy	0.89
Validation set ROC	0.84

Our algorithm exhibited a total accuracy of 89% at prediction of textile product recommendation based on customer sentiment. Its overall training accuracy was 91% with a validation training accuracy of 89% and a precision of 56% for negative sentiment recommendation and 96% for the positive sentiment recommendation. Hence, our algorithm is more efficient at prediction of positive sentiment as compared to negative sentiment according to the results in (Tables 2 and 3).

4 Results and Discussion

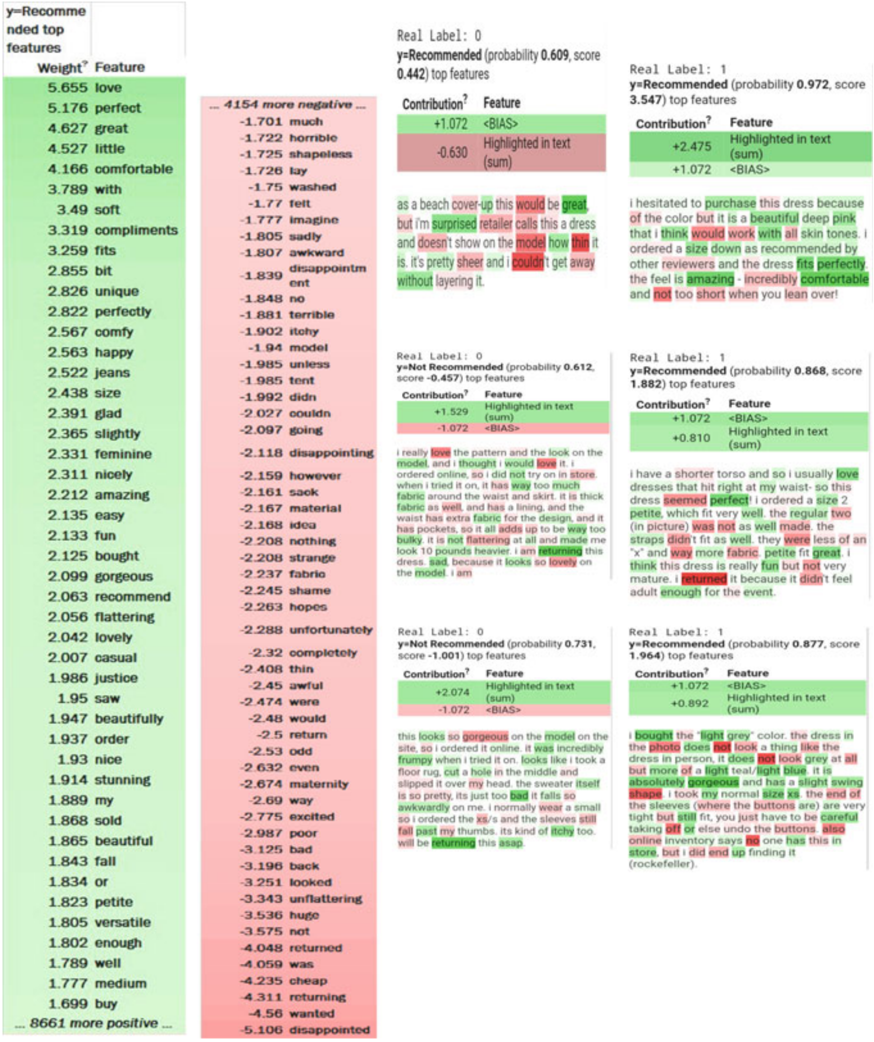
In this section, we illustrate and explain the simulated results obtained through the process of executing the methodology explained above for improved interpretation of customer sentiment.

The specific explanations of the various customer reviews (sentiment) depend on frequency and weights of the text (words) in the message. Depending on the scale of positivity or negativity represented in the "Specific explanations" column, the algorithm classified the reviews according to the weights of each feature exhibited in the text. The illustrations of the sentiment classifications are displayed below. The label is what was automatically assigned by the algorithm where (0—represents negative sentiment, i.e., reduced possibility of product recommendation based on customer experience, 1—represents positive sentiment, i.e., higher chances that the customer will recommend the product based on customer experience) as shown in (Table 4).

5 Conclusion

In this work, we proposed and demonstrated a reliable AI approach to that transparently analyzes textile customer sentiments. We predicted the textile product recommendation based on customer reviews and ratings, and we illustrated how textile product development and distribution can be optimized for a positive return on investment by utilizing a mix to emerging 4th Industrial Revolution technologies like Natural Language Processing and Explainable Artificial Intelligence. We also illustrated the benefits of Artificial Intelligence (AI) technology adoption for e-commerce and business production, development, and distribution operations.

Table 4 Interpretable recommendation results



y=Recommended top features

Weight?	Feature
5.655	love
5.176	perfect
4.627	great
4.527	little
4.166	comfortable
3.789	with
3.49	soft
3.319	compliments
3.259	fits
2.855	bit
2.826	unique
2.822	perfectly
2.567	comfy
2.563	happy
2.522	jeans
2.438	size
2.391	glad
2.365	slightly
2.331	feminine
2.311	nicely
2.212	amazing
2.135	easy
2.133	fun
2.125	bought
2.099	gorgeous
2.063	recommend
2.056	flattering
2.042	lovely
2.007	casual
1.986	justice
1.95	saw
1.947	beautifully
1.937	order
1.93	nice
1.914	stunning
1.889	my
1.868	sold
1.865	beautiful
1.843	fall
1.834	or
1.823	petite
1.805	versatile
1.802	enough
1.789	well
1.777	medium
1.699	buy
… 8661 more positive …	

Weight	Feature
… 4154 more negative …	
-1.701	much
-1.722	horrible
-1.725	shapeless
-1.726	lay
-1.75	washed
-1.77	felt
-1.777	imagine
-1.805	sadly
-1.807	awkward
-1.839	disappointment
-1.848	no
-1.881	terrible
-1.902	itchy
-1.94	model
-1.985	unless
-1.985	tent
-1.992	didn
-2.027	couldn
-2.097	going
-2.118	disappointing
-2.159	however
-2.161	sack
-2.167	material
-2.168	idea
-2.208	nothing
-2.208	strange
-2.237	fabric
-2.245	shame
-2.263	hopes
-2.288	unfortunately
-2.32	completely
-2.408	thin
-2.45	awful
-2.474	were
-2.48	would
-2.5	return
-2.53	odd
-2.632	even
-2.674	maternity
-2.69	way
-2.775	excited
-2.987	poor
-3.125	bad
-3.196	back
-3.251	looked
-3.343	unflattering
-3.536	huge
-3.575	not
-4.048	returned
-4.059	was
-4.235	cheap
-4.311	returning
-4.56	wanted
-5.106	disappointed

Real Label: 0
y=Recommended (probability **0.609**, score **0.442**) top features

Contribution?	Feature
+1.072	<BIAS>
-0.630	Highlighted in text (sum)

as a beach cover-up this would be great, but i'm surprised retailer calls this a dress and doesn't show on the model how thin it is. it's pretty sheer and i couldn't get away without layering it.

Real Label: 1
y=Recommended (probability **0.972**, score **3.547**) top features

Contribution?	Feature
+2.475	Highlighted in text (sum)
+1.072	<BIAS>

i hesitated to purchase this dress because of the color but it is a beautiful deep pink that i think would work with all skin tones. i ordered a size down as recommended by other reviewers and the dress fits perfectly. the feel is amazing - incredibly comfortable and not too short when you lean over!

Real Label: 0
y=Not Recommended (probability **0.612**, score **-0.457**) top features

Contribution?	Feature
+1.529	Highlighted in text (sum)
-1.072	<BIAS>

i really love the pattern and the look on the model, and i thought i would love it. i ordered online, so i did not try on in store. when i tried it on, it has way too much fabric around the waist and skirt. it is thick fabric as well, and has a lining, and the waist has extra fabric for the design, and it has pockets, so it all adds up to be way too bulky. it is not flattering at all and made me look 10 pounds heavier. i am returning this dress. sad, because it looks so lovely on the model. i am

Real Label: 1
y=Recommended (probability **0.868**, score **1.882**) top features

Contribution?	Feature
+1.072	<BIAS>
+0.810	Highlighted in text (sum)

i have a shorter torso and so i usually love dresses that hit right at my waist- so this dress seemed perfect! i ordered a size 2 petite, which fit very well. the regular two (in picture) was not as well made. the straps didn't fit as well. they were less of an "x" and way more fabric. petite fit great. i think this dress is really fun but not very mature. i returned it because it didn't feel adult enough for the event.

Real Label: 0
y=Not Recommended (probability **0.731**, score **-1.001**) top features

Contribution?	Feature
+2.074	Highlighted in text (sum)
-1.072	<BIAS>

this looks so gorgeous on the model on the site, so i ordered it online. it was incredibly frumpy when i tried it on. looks like i took a floor rug, cut a hole in the middle and slipped it over my head. the sweater itself is so pretty, its just too bad it falls so awkwardly on me. i normally wear a small so i ordered the xs/s and the sleeves still fall past my thumbs. its kind of itchy too. will be returning this asap.

Real Label: 1
y=Recommended (probability **0.877**, score **1.964**) top features

Contribution?	Feature
+1.072	<BIAS>
+0.892	Highlighted in text (sum)

i bought the "light grey" color. the dress in the photo does not look a thing like the dress in person, it does not look grey at all but more of a light teal/light blue. it is absolutely gorgeous and has a slight swing shape. i took my normal size xs. the end of the sleeves (where the buttons are) are very tight but still fit, you just have to be careful taking off or else undo the buttons. also online inventory says no one has this in store, but i did end up finding it (rockefeller).

Our future work will be focused on building a standardized sentiment analysis Application Programming Interface (API) for e-commerce businesses to explainably analyze and interpret the experiences of their customers through reviews and ratings. We also recommend that textile product development and distribution stakeholders embrace and adopt Artificial Intelligence to innovatively guide their product development for local and global markets based on customer preferences.

References

1. Miranda, M.D., Sassi, R.J.: Using sentiment analysis to assess customer satisfaction in an online job search company. In: Business Information Systems Workshops, pp. 17–27. Springer International Publishing, Cham (2014)
2. Marvin, G., Jackson, M., Alam, M.G.R.: A machine learning approach for employee retention prediction. In: 2021 IEEE Region 10 Symposium (TENSYMP) (2021)
3. [Online]. Available: https://doi.org/10.1093/icc/dtaa059-6. Accessed 28 Oct 2021
4. Lee, J., Lee, K.: Is the fourth industrial revolution a continuation of the third industrial revolution or something new under the sun? Analyzing technological regimes using US patent data. Ind. Corp. Chang. **30**(1), 137–159 (2021)
5. Bai, P., Xia, Y., Xia, Y.: Fusing knowledge and aspect sentiment for explainable recommendation. IEEE Access **8**, 137150–137160 (2020)
6. [Online]. Available: https://doi.org/10.3390/joitmc703019. Accessed 30 Oct 2021
7. Dwivedi, Y.K., et al.: Setting the future of digital and social media marketing research: perspectives and research propositions. Int. J. Inf. Manage. **59**(102168), 102168 (2021)
8. Davis, F.: Social Media Flagship Report, Q1 2019. Digitalmarketingcommunity.com. [Online]. Available: https://www.digitalmarketingcommunity.com/researches/social-flagship-report-q1-2019/. Accessed 04 Oct 2021
9. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
10. [Online]. Available: https://doi.org/10.1016/B978-1-78242-379-9.00007. Accessed 28 Oct 021
11. Michel, S., Kreuzer, M., Stringfellow, A., Schumann, J.H.: Mass-customised products: are they bought for uniqueness or to overcome problems with standard products? J. Cust. Behav. **8**(4), 307–327 (2009)
12. "Welcome to ELI5's documentation!—ELI5 0.11.0 documentation. Readthedocs.io. [Online]. Available: https://eli5.readthedocs.io/en/latest/. Accessed 29 Oct 2021
13. Chen, H.-Y., Lee, C.-H.: Vibration signals analysis by explainable artificial intelligence (XAI) approach: application on bearing faults diagnosis. IEEE Access **8**, 134246–134256 (2020)
14. Campers, H.: Eli5: Eli5 Notebook/Journal, Notebook/Journal. Independently Published (2019)
15. Agarap, A.F.M.: Statistical analysis on E-commerce reviews, with sentiment classification using bidirectional recurrent neural network. Arxiv.org. [Online]. Available: http://arxiv.org/abs/1805.03687v2. Accessed 28 Oct 2021
16. Loper , E., Bird, S.: NLTK: The Natural Language Toolkit. arXiv [cs.CL] (2002)

Automatic Multi-document Summarization for Bangla News Text Using a Novel Unsupervised Approach

Mohammad Hemayet Ullah, Maimuna Rahman, Bonosree Roy, and Md. Mohsin Uddin

Abstract Digitalization made the world remarkably agile; therefore, plenty of news is published on a related topic in different newspapers. An automatic text summarizer can contain all the primary information on a particular topic to understand the main essence of news or an article quickly and efficiently. We have developed a novel extractive-based multi-document summarizer for Bengali news. A very few summarizers have been introduced for the Bangla language. To the best of our knowledge, the word embedding technique has not been used in any of the prior summarizers. The selected sentences of the summaries' semantic and syntactic relationship are maintained through the utility of word embedding. In this paper, we have used predictive-based embedding, which is continuous bag-of-words-based word2vec algorithm. Since it is a multi-document summarizer, redundant sentences are present in the dataset and are removed by a clustering technique. The proposed model gives a statistically significant result in terms of ROUGE-L F1-measure, which outperforms the baseline system on the same scale. The proposed model achieves state-of-the-art performance and can do a better readable and informative synopsis, which might help the reader gain vital information reliably fast.

Keywords Multi-document summarization · Bengali news · Word2vec · CBOW · Extractive summary

1 Introduction

Today's world moves forward at a more incredible speed, and within the second, numerous information, news, and articles are on the Internet. When a person looks for a particular segment of information in the given space, it takes much more time to find out that information. In this case, a brief would be helpful. Summarizing

M. H. Ullah (✉) · M. Rahman · B. Roy · Md. M. Uddin
East West University, Dhaka, Bangladesh
e-mail: hemayet.nirjhoy@gmail.com

Md. M. Uddin
e-mail: mmuddin@ewubd.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_34

from the source can be time consuming for readers. An automatic summarizer could be the most effective solution to solve this hurdle because it will do the task more efficiently. Text summarizing is one of the most challenging tasks in the natural language processing domain. With the assistance of an automatic text summarizer, readers can get the information in a shorter time, and it is less biased than the summaries generated by humans.

Text summarization is the process which sums up the salient points of the bestowed document while containing all the information without any redundant sentences. It illustrates the principal notion of the initial record within half of the space. Summarizing a text is a difficult endeavor because it necessitates both cognitive ability and comprehension in order to develop summaries [16]. Two different approaches are being followed in natural language processing (NLP) to generate summaries: extractive summarization and abstractive summarization. As summarizing is a highly complex task, therefore to avoid difficulty, extractive summarization has been used. This technique works like a copy-paste fashion [16]. This method chooses a subset of essential words, phrases, or sentences extracted from the primary source, merges them, and summarizes where the inputs are taken directly from the text. Mainly, an extractive method does not need to rephrase the original text. While highlighting the main points, this method also avoids redundancy.

In contrast, an abstractive summarization method paraphrases the document and minimizes the original document while keeping the essential part. Very few works have been done for abstractive summarization because of its complex nature. While generating a summary from the source, the source can be two types: single document and multi-document summarization. The single document summarization deals with only one topic under one document, where multiple documents under one case are considered in the multi-document summarization. A multi-document summary succinctly portrays the information held in a group of records. They assist users in comprehending and displaying the document cluster in a consistent manner [17].

Automatic text summarization is a very well-known, critical, and challenging topic in the natural language processing domain. Previously in machine learning, many works have been done regarding this topic, mostly in English. Perhaps a small amount of work has been implemented in the Bengali language, but those systems are inefficient. A multi-document extractive Bangla news summarization system has been proposed to contribute to the Bangla language. The proposed system will contain a general extractive summary from multiple documents and also give a statistically significant result compared to baseline approach.

We have used predictive-based embedding, which is word2vec, in this research work. The reason behind choosing the word2vec is as follows, this method was not used in the previous researchers. Moreover, it can put together similar terms in the vector space and assume the next or previous word according to the context. Therefore, we have built a model with a word2vec algorithm. Among two techniques of word2vec, the continuous bag-of-words (CBOW) model to implement the system, which is selected by researches follows the extractive method to generate summaries. Latent semantic analysis (LSA), a baseline method, is also implemented for comparison with the proposed method.

We have provided a brief review of the previous studies of multi-document summarization in Sect. 2. The preparation of the dataset and the CBOW models implementation is discussed in Sect. 3. In Sect. 4, the experimental analysis, along with the result, is mentioned. Finally, in the last Sect. 6, we concluded with the future vision of our research.

2 Literature Review

The researchers had worked on different methods to deploy an efficient summarizer for the Bangla language [17]. Bangla is the most used language, and it is the seventh most spoken language in the world [9]. Due to the Bangla language text's multi-faceted structure, maximum work in text summarization is conducted in English. This research has implemented the extraction-based summarizer based on word2vector's CBOW model, a shallow neural network to efficiently learn word embedding in a given corpus text. The authors in [17] used the famous data mining method, the term frequency (TF), for generating summaries from multiple documents. However, they have incorporated the frequency count with some statistical tasks to score the words and maintain relevancy in the given corpus sentences. In the research [2], a new model is based on a simple extractive summarization method. The authors implemented an advanced sentence scoring method which are different associations and linguistic rules. There are two steps summarization techniques in their work where prime sentence selection is the vital one. The topic of the texts is obtained utilizing the word and sentence frequency. Another work [9] developed a hybrid summarizing system for Bengali documents. Where the system formulated complete and neutral summaries based on the combination of three methods. The methods are scoring the keyword, analyzing the sentiment of the sentences, and the interrelation of sentences. They also consider time analysis to determine the specific time for generating a single summary. In [8], an automatic Bangla summarizer was built using statistical and mathematical procedures, including grammatical rules to generate a precise summary. Here, a primary summary is generated at first using the sentence score and position technique. After analyzing the prior summary and the help of Bangla grammatical rules and the sentence simplification method, the final summary is created in this paper [8]. Here, the evaluation was made for a single document with the manual summaries written by humans and the system-generated summaries. A study [14] presented a centroid-based summarizer named MEAD for summarizing multiple documents. It automatically summarizes news reports by using cluster centroids. The cluster centroids are produced by detecting the topic and the tracking system. This system determines the most probable sentences relevant to the cluster topic utilizing the information of the centroid. Here, a unique utility-based technique, named cluster-based sentence utility (CBSU), has been used for evaluation purposes. Another study [3] suggested an A* algorithm to determine the best score summaries from multi-document using the method to construct summaries for English at a certain length. This system generates extractive summaries from a list of ranked sentences and can learn feature weight. A discriminative training algorithm is also proposed to maximize the quality of a summary. Since multi-document summarization needs multi-objectives, the researchers optimized the artificial bee colony algorithm to multi-objective artificial bee colony (MOABC). The algorithm is designed to increase the content's coverage and decrease the redundancy from a collection of documents for the English language. Some researchers used a multi-document summarizer which uses the deep neural network's representational ability and graphs for sentence connectedness. They developed a graph convolutional network (GCN) architecture employed to a personalized discourse graph, where three-sentence relations graphs have been considered in their approach. The combination of sentence relations and graphs gives a better result than those models which do not use graphs [18]. In paper [5], an abstractive summary work has been performed by a neural attention approach. Their model facilitates an LSTM networking model in recognition with encoder–decoder and generates long string of words with coherent, human-like sentences including the information from the source material. Another work for multi-document summarizer [1] used the continuous bag of words for the Arabic language to preserve the original information. Another study [13], using the distributed bag of words, they introduce a multi-document summarizer. Their approach is an unsupervised document-level reconstruction framework based on centroid, which chooses summary sentences to decrease the reconstruction error. A unique sentence scoring feature which is a graph-based method is implemented for the Bangla news documents to generate a summary in [6]. This paper [7] had explored the quantity of research work on Bangla text summarization to draw the attention of the following researchers and provide them a clear direction in this arena. Fourteen approaches had been briefly explained with the advantages and disadvantages, including some opportunities for improvement, and a comparison has been made based on the incorporated features and evaluation outcome. In paper [11], latent semantic analysis is used for the single Bangla document text summarization. The researchers use LSA for finding semantically equivalent sentences. When comparing closely similar sentences with no common words, a bag-of-words representation fails to understand the semantic relationships between concepts. A novel and straightforward extractive approach were suggested in [16] to address this issue. It is based on the geometric sense of the centroid vector of a (multi) paper by leveraging the compositional properties of word embeddings. The structure of our system is motivated by [16], where a centroid-based summarization method was used for multi-document and multilingual single documents. However, our approach is different. In our work,

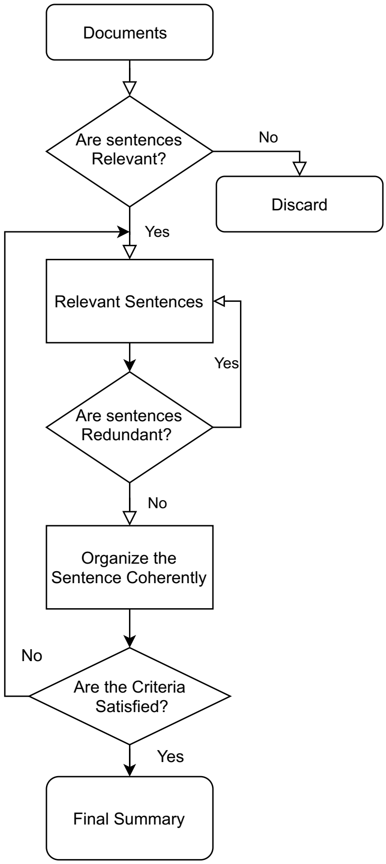
- We have developed a multi-document summarizer in the Bangla language utilizing the vector representation of word embedding.
- A clustering technique which is a connected component algorithm is used to remove similar sentences from the dataset.

- Several Bangla words have been collected from a well-known newspaper archive to train the word2vec's CBOW model. Here, it is an unsupervised method. Therefore, labeled data is not required for this purpose.

3 Proposed Method

The process of text summarization is very complex; on top of that, the Bangla language's structural pattern is more intricate. So to get a beautiful and elegant summary, this whole process has to go through many steps. Those steps of our work are described in the following sections below including an illustration 1.

Fig. 1 Proposed method



3.1 Overview of the Dataset

Collection of the Data The proposed method is unsupervised; on this account, labeled data is not required for training the datasets. Moreover, numerous Bangla words are collected from Prothom-Alo[1] archived for training purposes. To create the dataset, we have collected Bangla news reports from some of the leading newspapers in Bangladesh, such as Prothom-Alo, Kaler Kantho,[2] and Jugantor.[3]

Dataset Preparation A set consisting of four news reports from each newspaper based on the same event and almost the same headline in the dataset. Several reference summaries are produced from the set of data by different participants, and the best one is taken for the final reference summaries, which includes the evaluation of human judges. Though there is no judgment criteria for summary as humans have the distinct inclination. Therefore, more informative and readable summaries are identified as a higher priority, and then, the dataset was finalized. Our goal is to summarize all the news in a single document and create precise, enlightening, and unbiased summaries, which are the three challenging tasks. The final summaries must be relevant and coherent without having any redundant information.

3.2 Preprocessing

Data preprocessing is an essential part of text summarization to make the text more understandable. It makes the data more predictable and also helps to analyze the task. Most of the time, the collected data contains unnecessary words and symbols which need to be removed to make a sophisticated model. The quality of the data improves to build a more reliable machine learning model. In the next part, we have described our preprocessing steps for our proposed work.

Training the Word2Vec Model We train the word2vec model with numerous data. Those data are collected from the Prothom-Alo archived that has been uploaded in Kaggle.[4] From 2013 to 2019, we randomly selected two to ten thousand news documents published by the daily newspaper Prothom-Alo. These documents were used to train different word2vec models. After examining the trained model, we selected those models, which gives a better mapping between words. The collected news articles are divided into sentences and then from sentences into words. Many Bangla words train the model and the Gensim word embedding library[5] for the training word2vec model.

---

[1] https://www.prothomalo.com/.

[2] https://www.kalerkantho.com/.

[3] https://www.jugantor.com/.

[4] https://www.kaggle.com/.

[5] https://radimrehurek.com/gensim/models/word2vec.html.

Stop Word Deduction Every language has some words which do not carry any meaning in the sentences. Those words are known as stop words. To analyze and build the model, stop words are deducted from the texts. It is a prevalent task that is done in the preprocessing phase. The Bangla language has many words frequently used in the context, though none have thematic meaning. For improving the accuracy, stop words are not considered in the case of text processing, while keywords get the priority instead. However, the deduction of stop words does not affect the meaning of the sentence. In the perspective of the Bangla text, there is no authorized list of stop words. Therefore, a list of four hundred stop words collected and ignored all those listed words from every sentence of documents.

To produce an appropriate, concise, and meaningful summary, it is the most complicated work in text summarization. Since our system summarizes multiple documents in a single document, similar sentences will emerge in the final summary. On the other hand, the dataset is based on similar events so that the system might extract irrelevant sentences, or the sentences, in summary, might not be serialized according to the period. Therefore, maintaining relevancy as well as redundancy and coherency is the main task for producing a more reliable summary.

Relevancy The most crucial task for summarization is maintaining disambiguation. The system should extract the most relevant information for a more positive perception of the event or topic that will summarize. Relevancy is met based on similarities between newspaper headlines and sentences. Similarity scores have been measured for every sentence of each document with a corresponding headline and sorted all those sentences based on the similarity scores and selected the most relevant sentences. Because word2vec measures the semantic similarity between words that ensure the most relevant information, all those sentences of every document are chosen and merged and then sent to the next phase 1.

Algorithm 1 Relevancy

procedure RELEVANT- SENTENCES($D[d_1, d_22, \ldots, d_n], h$) ▷ where, D is the set of documents to summarize and h is the title of documents
    $S[s_1, s_2, ..., s_n] \leftarrow$ *docToSent(D)*
    for each $s\ in\ S$ **do**
        relevancy $\leftarrow word2vecModel(s, h)$
        if relevancy < threshold **then**
            $continue$
        else
            $R[r_1, r_2, \cdots, r_p] \leftarrow s$
    return $set\ of\ relevant\ sentences\ R$

Redundancy Redundant information deduction is another vital task for making the summary engaging. In summary, the redundant information is evident as the previous phase merged all the sentences in one document, $R[r_1, r_2, \ldots, r_p]$. The clustering technique has been used for deducting redundant sentences for the final draft. Firstly, similarities between every sentence have been measured and clustered in different sets, where a set of similar sentences now belongs together. Few sets are units, and few have the same components as we measured the similarity from all ends. The connected component algorithm is used to ensure the unique presence of each sentence and clustered them all sentences where no duplicate sentences arise, neither in clustered set and unit set. Again, the most relevant sentences are chosen as Q from each cluster and united for summary until the word limit exceeds the summary draft 2. These are the final sentences chosen for a multi-document summary.

Algorithm 2 Redundancy

procedure REDUNDANT- SENTENCES($R[r_1, r_2, \cdots, r_p]$) ▹ where, R is the set of most relevant sentences
    for each r_i *in* R **do**
        for each r_j *in* R **do**
            $S[i][j] \leftarrow Similarity(r_i, r_j)$ ▹ $where,\ r_i \neq r_j$
    for each row or i of S **do**
        $U[u_1, u_2, \cdots, u_n] \leftarrow SimilarSet(S[i])$ ▹ U belongs to a universal set having common element in multiple sets
    $U'[u'_1, u'_2, \cdots, u'_n] \leftarrow ConnectedComponent(U)$ ▹ No set has common attribute at all.
    $Q[q_1, q_2, \cdots, q_n] \leftarrow SelectSentences(U', w)$ ▹ w denotes the word limit, and f for final selection
    return $final\ sentences$ **Q**

Coherency Coherence is the foremost thing in a summary to make the summary more smooth and logical. After checking the redundancy, the stack of sentences $Q[q_1, q_2, \ldots, q_f]$ is ready to finalize a summary F which consists of multiple documents. However, the sentences are not in the correct order to understand the summary. The original position of the sentences in the document is taken into account for resolving this problem; if more than one sentences have the same position, then their relevance score has been considered to finalize their position. The final script is re-tested to ensure the desired quality and word limit of the summary. This inspection is a comparison with the information which have been deducted in a redundant phrase. The most relevant sentences are selected while prioritizing the unit sets.

4 Implementation of Baseline

Latent semantic analysis (LSA) [10] is a standard method for summarization; for this reason, we have taken LSA as a baseline in this work. This method represents the contextual meaning of the text through statistical analysis. Based on the term frequency(TF) [15] score of each sentence of all documents, the initial matrix has been generated, excluding the stop words as the stop words have been ignored during frequency calculations. Therefore, the inverse document frequency (IDF) [15] did not calculate. The initial one becomes a sparse matrix having a large dimension depending on the size and number of the document provided for summarization. For dimensionality reduction, singular value decomposition (SVD) [4] (version of principal component analysis [18]) is applied. The final matrix has contextually highlighted the relevant information. In terms of multi-document summarization, it collects redundant information multiple times that makes the summary very inadequate to publish. Based on the proposed work and experience, redundant checking has been made, especially for this multi-document summary implementation. Hence, LSA generates a comparatively better summary than the previous one. Set of documents $D = [d_1, d_2, \ldots, d_m]$ have been transformed into sentences $S = [s_1, s_2, \ldots, s_n]$. In matrix A, the $rows(m)$ and $columns(n)$ have been initialized by $zero$ where $m =$ number of documents and $n =$ number of sentences. The frequency of each word fw_i of a sentence s_i counted in every document. Summing up each word score of a sentence gives the frequency of a sentence's fs_i. That frequency is divided by the total terms contained by each document. SVD (A) gives a set of matrixes $[U(m \times n), S(n \times n), V^T(n \times n)]$, where U and V^T both are orthogonal matrix, S is a diagonal matrix, and $(m \times n)$ represents the row and column of each matrix, respectively. We have a new set of matrix $[U'(m \times k), S'(k \times k), V'(n \times k)]$ after selecting k element from the previous matrix $[U(m \times n), S(n \times n), V^T(n \times n)]$. Reconstructing a new set of the matrix gives a new $(m \times n)$ matrix A'. Final sentences are selected from the reconstructed matrix A' avoiding redundancy. In addition, this redundant technique is applied for improving the generic LSA method for multi-document summarization. Otherwise, extracted sentences carry repeated information.

5 Result Analysis

To evaluate the summaries, we have used ROUGE (recall-oriented understudy) [12] matrices for the gisting evaluation for both the proposed work and the baseline method. ROUGE-N, ROUGE-S, and ROUGE-L measure the amount of detail in texts to compare with system and reference summaries.

ROUGE-N measures the overlap of N-grams between the system and reference summaries. The most used ROUGE-1 and ROUGE-2 refer to the overlap of unigram which means each word and bigram in both summaries. On the other hand, ROUGE-

L is based on the longest common subsequence statistics [9, 11]. The frequently used metrics of ROUGE [12] are *recall*, *precision*, and *F-score* which can be described as the following formula:

$$\text{Recal} = \frac{\text{number of overlapped words}}{\text{total words in reference summary}}$$

$$\text{Precission} = \frac{\text{number of overlapped words}}{\text{total words in system-generated summary}}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

According to the ROUGE evaluation, both Word2Vec and LSA generated summaries give a satisfactory result exhibited in Table 1

Here, Fig. 2 is the reference summary which is done by humans. Figure 4 is the baseline summary, and Fig. 3 is the summary of our proposed system. Our system is novel because it is relevant to the topic, has no redundant sentence, and maintains coherency. On the other hand, several redundant sentences in the baseline summary appear in four and five number lines. There is also a punctuation error in the baseline summary (Figs. 3 and 4).

Table 1 ROUGE score comparison

Summary	Proposed method			Baseline method		
	F-score	Precision	Recall	F-score	Precision	Recall
ROUGE-1	0.61	0.55	0.67	0.42	0.44	0.42
ROUGE-2	0.37	0.32	0.44	0.18	0.17	0.19
ROUGE-L	0.43	0.39	0.47	0.26	0.27	0.26

আবরার হত্যা মামলার নথি উপস্থাপন করা হয় ঢাকার অতিরিক্ত মূখ্য মহানগর হাকিম কায়সারুল ইসলামের আদালতে। গত ১৩নভেম্বর বাংলাদেশ প্রকৌশলী বিশ্ববিদ্যালয়ের (বুয়েট) শিক্ষার্থী আবরার হত্যা কান্ডের ঘটনায় ২৫জনকে আসামী করে চার্জশীট গ্রহণ ও পলাতক মোর্শেদুর জামান জিসান, এহতেশামুল রাব্বি তানিম, মোর্শেদ অমর্ত্য ইসলাম ও মোস্তফা রাফিদ নামে ৪আসামীর বিরুদ্ধে গ্রেপ্তারি পরোয়ানা জারি করার মাধ্যমে আবরার হত্যার বিচার কার্যক্রম পুনঃরায় শুরু হয়েছে। ফাহাদের বাবা বরকত উল্লাহ ১৯জনকে আসামি করে চকবাজার থানায় হত্যা মামলা দায়ের করেন এবং অভিযুক্ত ২৫ আসামির মধ্যে ২১জনকে গ্রেফতার করা হয়েছে। এই মামলায় স্বীকারোক্তি মূলক জবানবন্দি দিয়েছে ৮ আসামি।

Fig. 2 Reference summary

বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়ের (বুয়েট) শিক্ষার্থী আবরার ফাহাদকে পিটিয়ে হত্যার ঘটনায় ২৫ জনকে আসামি করে করা চার্জশিট গ্রহণ করেছেন আদালত। একই সঙ্গে পলাতক চার আসামির বিরুদ্ধে গ্রেপ্তারি পরোয়ানা জারি করেছেন। এর মধ্য দিয়ে আবরার হত্যার বিচার কার্যক্রম শুরু হলো। আজ সকালে আবরার হত্যা মামলার নথি উপস্থাপন করা হয় ঢাকার অতিরিক্ত মুখ্য মহানগর হাকিম কায়সারুল ইসলামের আদালতে। গত ১৩ নভেম্বর আবরার হত্যা মামলায় ২৫ জনকে আসামি করে চার্জশিট দেওয়া হয়। যাদের বিরুদ্ধে পরোয়ারা জারি করা হয়েছে তারা হলেন-মোর্শেদুজ্জামান জিসান, এহতেশামুল রাব্বি তানিম, মোর্শেদ অমত্য ইসলাম ও মোস্তবা রাফিদ। এ ঘটনায় ফাহাদের বাবা মো. বরকত উল্লাহ ১৯ জনকে আসামি করে চকবাজার থানায় একটি হত্যা মামলা দায়ের করেন। অভিযোগপত্রভুক্ত ২৫ আসামির মধ্যে ২১ জনকে ইতিমধ্যে গ্রেপ্তার করা হয়েছে। এখন পর্যন্ত পলাতক রয়েছেন জিসান, তানিম, মোরশেদ, মোশতুবা রাফি। এ মামলায় স্বীকারোক্তিমূলক জবানবন্দি দিয়েছেন আট আসামি।

Fig. 3 Proposed method generated summary

বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়ের (বুয়েট) ছাত্র আবরার ফাহাদ হত্যা মামলার অভিযোগপত্র গ্রহণ করেছেন আদালত। একই সঙ্গে এই মামলার পলাতক চার আসামির বিরুদ্ধে গ্রেপ্তারি পরোয়ানা জারি করা হয়েছে। ২১ জনের মধ্যে ১৬ জনের নাম আবরারের বাবার করা হত্যা মামলার এজাহারে আছে। পরে জানা যায়, বিশ্ববিদ্যালয়ের ছাত্রলীগের নেতা-কর্মীরা তাঁকে পিটিয়ে হত্যা করেছেন। যাদের বিরুদ্ধে পরোয়ারা জারি করা হয়েছে তারা হলেন-মোর্শেদুজ্জামান জিসান, এহতেশামুল রাব্বি তানিম, মোর্শেদ অমত্য ইসলাম ও মোস্তবা রাফিদ। তাদের বিরুদ্ধে আজ গ্রেফতারি পরোয়ানা জারি করা হলোবাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়ের (বুয়েট) শিক্ষার্থী আবরার ফাহাদকে পিটিয়ে হত্যার ঘটনায় ২৫ জনকে আসামি করে করা চার্জশিট গ্রহণ করেছেন আদালত। সোমবার ঢাকার অতিরিক্ত মহানগর হাকিম কায়সারুল ইসলাম মামলার অভিযোগপত্র (চার্জশিট) গ্রহণ করেন।

Fig. 4 Baseline method generated summary

6 Conclusion and Future Work

In this paper, we present an efficient way to summarize multi-document Bangla news documents. The system follows the extractive method to make the final summary. Moreover, our approach exploits the simplicity of word2vec's continuous bag-of-words (CBOW) technique. A baseline system using LSA is also developed to highlight the effectiveness of the proposed summary in our research. However, in future, the goal is to develop a summarizer that will create abstractive summaries as well as to generalize the news according to relevant categories.

References

1. Abdulateef, S., Khan, N.A., Chen, B., Shang, X.: Multidocument Arabic text summarization based on clustering and word2vec to reduce redundancy. Information **11**(2), 59 (2020)
2. Abujar, S., Hasan, M., Shahin, M., Hossain, S.A.: A heuristic approach of text summarization for bengali documentation. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–8. IEEE (2017)
3. Akter, S., Asa, A.S., Uddin, M.P., Hossain, M.D., Roy, S.K., Afjal, M.I.: An extractive text summarization technique for Bengali document (s) using k-means clustering algorithm. In: 2017 IEEE International Conference on Imaging, Vision and Pattern Recognition (icIVPR), pp. 1–6. IEEE (2017)
4. Andrews, H., Patterson, C.: Singular value decomposition (SVD) image coding. IEEE Trans. Commun. **24**(4), 425–432 (1976)

5. Bhattacharjee, P., Mallick, A., Islam, M.S., et al.: Bengali abstractive news summarization (bans): a neural attention approach. In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering, pp. 41–51. Springer (2021)
6. Ghosh, P.P., Shahariar, R., Khan, M.A.H.: A rule based extractive text summarization technique for Bangla news documents. Int. J. Mod. Educ. Comput. Sci. **11**(12), 44 (2018)
7. Haque, M.M., Pervin, S., Hossain, A., Begum, Z.: Approaches and trends of automatic Bangla text summarization: challenges and opportunities. Int. J. Technol. Diffus. (IJTD) **11**(4), 67–83 (2020)
8. Hossain, M.M.: Automatic text summarization for Bengali language including grammatical analysis. Int. J. Sci. Technol. Res. **8**, 288–292 (2019)
9. Islam, M., Majumdar, F.N., Galib, A., Hoque, M.M.: Hybrid text summarizer for Bangla document (2020)
10. Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, L.R.G., Matthews, C., Lamb, R.: Developing summarization skills through the use of lsa-based feedback. Interact. Learn. Environ. **8**(2), 87–109 (2000)
11. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp. 150–157 (2003)
12. Lin, C.Y., Och, F.: Looking for a few good metrics: Rouge and its evaluation. In: Ntcir Workshop (2004)
13. Mani, K., Verma, I., Meisheri, H., Dey, L.: Multi-document summarization using distributed bag-of-words model. In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 672–675. IEEE (2018)
14. Radev, D.R., Jing, H., Styś, M., Tam, D.: Centroid-based summarization of multiple documents. Inf. Proc. Manage. **40**(6), 919–938 (2004)
15. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the First Instructional Conference on Machine Learning, vol. 242, pp. 29–48. Citeseer (2003)
16. Rossiello, G., Basile, P., Semeraro, G.: Centroid-based text summarization through compositionality of word embeddings. In: Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, pp. 12–21 (2017)
17. Uddin, A., Sultana, K.Z., Alam, M.: A multi-document text summarization for Bengali language (2019)
18. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometr. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)

Toward Embedding Hyperparameters Optimization: Analyzing Their Impacts on Deep Leaning-Based Text Classification

Md. Rajib Hossain and Mohammed Moshiul Hoque

Abstract In the last few years, an enormous amount of unstructured text documents has been added to the World Wide Web because of the availability of electronics gadgets and increases the usability of the Internet. Using text classification, this large amount of texts are appropriately organized, searched, and manipulated by the high resource language (e.g., English). Nevertheless, till now, it is a so-called issue for low-resource languages (like Bengali). There is no usable research and has conducted on Bengali text classification owing to the lack of standard corpora, shortage of *hyperparameters tuning* method of text embeddings and insufficiency of embedding model evaluations system (e.g., intrinsic and extrinsic). Text classification performance depends on *embedding features*, and the best *embedding hyperparameter* settings can produce the best embedding feature. The embedding model default hyperparameters values are developed for high resource language, and these hyperparameters settings are not well performed for low-resource languages. The *low-resource hyperparameters tuning* is a crucial task for the text classification domain. This study investigates the influence of embedding hyperparameters on Bengali text classification. The empirical analysis concludes that an automatic embedding hyperparameter tuning (AEHT) with convolutional neural networks (CNNs) attained the maximum text classification accuracy of 95.16 and 86.41% for BARD and IndicNLP datasets.

Keywords Natural language processing · Low-resource text classification · Hyperparameters tuning · Embedding · Feature extraction

1 Introduction

Text classification is a fundamental research problem for natural language processing (NLP), where each document is assigned a tag or labeled based on text semantics. In recent times, low-resource text classification has been challenging due to the shortage

Md. R. Hossain · M. M. Hoque (✉)
Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh
e-mail: moshiul_240@cuet.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_35

of linguistic resources and proper research guidelines. Bengali is the seven most widely spoken language in the world, and two hundred forty-five million people in Bangladesh and some parts of India speak this language [1]. In the age of *industrial evaluation*, lots of unstructured and unlabeled Bengali texts document have been added to the World Wide Web, and in the near future, this unlabeled text document will be the source of many historical events [2]. There are numerous applications of Bengali text classification such as document organization, news filtering, spam detection, opinion mining, socially harassment text identification, and patient report classification. On the other hand, Bengali text classification is also a pre-requisite task of machine translation, language identification, and word seance disambiguation. In this situation, Bengali text classification should be a sensible solution that organizes, searches, and manipulates a large number of unorganised text documents which reduces the cost, time, and meet-up of the pre-requisite NLP downstream tasks .

However, there are much research which have conducted on Bengali text classification based on statistical learning [3], deep learning [4] with pre-trained embedding model, and manual hyperparameters tuning method [5]. However, single research has been conducted on embedding model hyperparameters optimization, e.g., automatic embedding parameters identification (EPI) that can tune the hyperparameters for Bengali embedding models [1]. Nevertheless, that research does not explain the initial, manual, and automatic hyperparameters impact on Bengali text classification. In this study, we empirically show the impact of embedding hyperparameters on Bengali text classification and summarized the performance using two Bengali benchmarks datasets, e.g., Bengali article classification dataset (BARD) [6] and IndicNLP [7].

The major contributions and findings of this research are as follows:

- Investigate the embedding model hyperparameters impact on Bengali text classification.
- Evaluate the embedding model performance using two Bengali benchmarks corpora, e.g., BARD and IndicNLP.
- Investigate the Bengali language diversity on embedding and classification models and provide future research direction.

2 Related Work

There is much research which has been conducted on low-resource text classification, including Bengali. However, most of the Bengali text classification studies have been conducted on pre-trained embedding model with non-tuned classification frameworks [6]. There are three embedding models which have been developed for word to vector representation such as Glove [8], FastText [9], and Word2Vec [10]. These embedding parameters are well-tuned for the English language, not for low-resource languages. As a result, the pre-tuned hyperparameters are not achieved a better result for the Bengali language due to large morphological variation between English and Bengali languages [1]. Hossain et al. [3] developed a statistical (e.g., SGD classifier)

Bengali text classification system using pre-tuned embedding hyperparameters and achieved 93.33% accuracy on self-build Bengali classification datasets. This model was not achieved to better accuracy because of statistical classifiers as well as not turning the embedding hyperparameters. Alam et al. [6] developed a Bengali article classification dataset (BARD) and classification performance measured using the statistical method with Word2Vec embedding model. However, this system calculates the folds wise performance, not considering the whole dataset performance. Word2Vec with BiLSTM-based method has developed by Humaira et al. [11] for Bengali text classification and maximum accuracy of 95.97% achieved for BARD dataset. But this system only considered the 10.0% data for validation purposes.

The CNN classifier with the FastText embedding model, Bengali text classification system, has achieved an accuracy of 96.85% for self-build datasets [4] with six different categories. This system builds the embedding model using pre-tuned embedding hyperparameters. A VDCNN and GloVe with embedding houseparents optimization-based Bengali text classification system have achieved maximum accuracy of 96.96% for a self-build dataset with 13 distinct categories. But the proposed system was evaluated using only one self-build dataset [1]. Rahman et al. [12] developed a Bengali text classification system using CNN and LSTM with character level embedding and attained the maximum accuracy of 92.41% for BARD datasets. They used only a 20% data sample for training and testing purposes and character level embedding not containing the sentence level semantic/syntactic features. However, this study provides an empirical summary of text classification and the impact of embedding model hyperparameters using two Bengali benchmarks datasets (e.g., BARD & IndicNLP).

3 Empirical Model

The automatic embedding hyperparameter tuning (AEHT) and CNN combinedly, named (AEHT-CNN)-based text classification framework, comprise three modules: automatic embedding hyperparameters tuning, text classification training, and testing. Figure 1 depicts the overview of the empirical framework. The embedding hyperparameters tuning module tuned and generated the embedding model, whereas the text classification model produces the classification model and testing module used for text classification. The following subsections describe the details of each module.

3.1 Automatic Embedding Hyperparameters Tuning

The embedding hyperparameters tuning module takes the input as embedding corpus (EC) and output the embedding model using embedding parameters identification (EPI) algorithm. The embedding hyperparameters tuning approach follow the philosophy of Witt et al. [13] research, and EPI algorithm is used for tuning hyperparameters

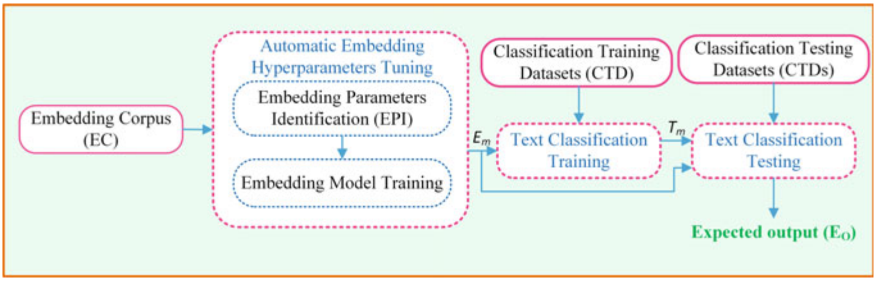
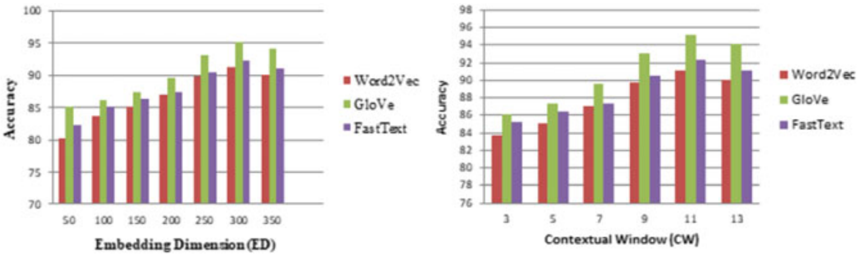


Fig. 1 AEHT-CNNs-based Bengali text classification framework



(a) Impact of embedding dimension (ED) on classification accuracy

(b) Impact of contextual windows (CW) on classification accuracy

Fig. 2 Two hyperparameters (e.g., CW & ED) impact on Bengali text classification accuracy using BARD datasets

[1]. In this research, we only considered the two most influential hyperparameters, e.g., embedding dimension (ED) and contextual windows (CW), for three embedding techniques (e.g., GloVe, FastText, and Word2Vec). The embedding hyperparameters are tuned using Eq. 1.

$$\Theta_t = F(EC, \Theta_i, \lambda, \Gamma) \tag{1}$$

Here, Θ_i and Θ_t indicate the initial and tuned embedding hyperparameters where Θ_i & $\Theta_t \in \{CW, ED\}$. F represents the hyperparameters tuning function, and λ denotes the intrinsic evaluation method as well as objective function of F. The Γ indicates that the total number of method is used in the tuning function and here, $\Gamma \in \{\text{Word2Vec, GloVe \& FastText}\}$. The embedding corpus ($EC$) and tuning function F have taken from the previous Bengali text classification research [1]. Figure 2 shows the overall performance of BARD datasets with two embedding hyperparameters (e.g., ED & CW).

Figure 2a, b shows the embedding dimension and contextual windows impact on classification task. The acronym ED indicates the embedding dimension, e.g., features for a word. The higher or lower ED does not mean much better correlated feature for embedding model. The ED and CW depend on the unlabeled embedding corpus morphological variation and amount of unique word on the corpus. Thus, this empirical visualization shows the outcome of the EPI algorithm impact on ED. For a similar reason, the contextual window CW varies from corpus to corpus for any low-resource language. As a result, Fig. 2b visualizes the CW's impact on the developed corpus. The EPI algorithm achieved the maximum accuracies of 91.18, 95.16, and 92.39% for Word2Vec+AEHT-CNNs, GloVe+AEHT-CNNs, and FastText+AEHT-CNNs using embedding dimension 300 with contextual windows 9. This embedding model is used for text to feature representation purpose in the training and testing modules.

3.2 Text Classification Training

The text classification training module takes the inputs as tuned embedding model (E_m) and classification training datasets (CTD). The CTD contains BARD and IndicNLP datasets. The output is the two classification models (T_m). The CNN classification method sequentially produces a classification model for the BARD dataset and IndicNLP dataset. The ith dataset jth sample text extracts the feature using Eq. 2.

$$I_f = M(E_m, CTD[i:j])_{j=1}^{|CTD[i:j]|} \tag{2}$$

Here, M denotes the mapping function where each word of $CTD[i:j]$ extract a feature vector from E_m and finally produce a 2D input feature matrix I_f. The $|CTD[i:j]|$ indicates the length of jth text. The feature matrix fed to the single layer multi-kernel CNNs models and the CNNs frameworks sequentially trained the model using Eq. 3 with Table 1 hyperparameters [5].

$$CF[k\text{th}:] = \sum_{i=1}^{r} I_f[i:\text{end}] \otimes F[k\text{th}:\text{end}] + B_i \tag{3}$$

Here, k^{th} kernel ($F[k^{\text{th}}:\text{end}]$) produces a $CF[k\text{th}:]$ single dimension feature vector, and r denotes the input feature matrix rows. The bias value B_i is added with the output convolution value. Now, the activation, pooling, and dropout operations are

Table 1 Best performing CNNs hyperparameters

Hyperparameters	Value
Batch size	128
Embedding	AEHT-GloVe
Embedding dimension	300
Filter size	2, 3, 4
No. of filters	128
Pooling size	2,3,4
Activation	LeakyReLU

sequentially applied to the feature vector (CF). Finally, the flattened feature vector is predicted the expected category name using softmax operation.

The CNNs model has produced two classification models for two different datasets (e.g., BARD & IndicNLP). These classification models are sued in the text classification testing module.

3.3 Text Classification Testing

The text classification module takes two inputs, e.g., text classification model T_m and classification testing datasets (CTDs), as well as output is expected category vector (E_O). The CTDs contain two datasets as BARD and IndicNLP. The ith dataset jth sample extracts the feature using Eq. 2 and prepared a input feature matrix I_{f_j} and calculate the expected category name using softmax Eq. 4.

$$E_O[1:N] = \max\left(\frac{e^{\Theta_{Wi} \times I_{f_j}}}{\sum_{k=1}^{k=|C|} e^{\Theta_{Wi}[i:k] \times I_{f_j}}}\right)_{l=1}^{N} \tag{4}$$

Here, Θ_{Wi} and $|C|$ represent the ith datasets classification model and total number of categories. N indicates the total number of the test sample in the ith datasets. For each dataset, try to assign an appropriate labeled using Eq. 4. The output vector $E_O[1:N]$ is used for performance measured purposes.

4 Experiments

The embedding hyperparameters tuning and CNN architecture have been deployed in the TensorFlow (V2) of Python 3.6, Scikit-learn, Gensim 4.1.2, and GloVe. The AEHT-CNNs architecture deployed at a core-i7 processor with NVIDIA GTX 1070 GPU 8 GB memory and physical memory of 32 GB.

4.1 Datasets

In this analysis, 100 semantic and syntactic word pairs datasets have been used for intrinsic or embedding model evaluation purposes [14]. The semantic word pair Kappa score is 78.00%, and the syntactic word pair Kappa score is 81.00%. At the same time, two types of corpora have been used for embedding and classification model generation purposes. The embedding corpus has been taken from the previous text classification research [1], and overall statistics show in Table 2.

Table 2 Embedding corpus (EC) statistics

Embedding corpus (EC) attributes	Attributes value
Number of text documents	969,000
Number of sentence	1,744,200
Maximum number of sentence (in a text)	170
Minimum number of sentence	1
Average number of sentence per document	7
Number of words	200,081,093
Number of unique words	9,517,390
Estimated required memory	16.7 GB

Table 3 Classification datasets statistics

Corpus attribute	BARD	IndicNLP
No. of category	5	6
No. of training texts	263,299	11,300
No. of testing texts	112,921	2,826
Max sentence (in a text)	121	89

This is the most extensive Bengali embedding corpus in terms of unique words (e.g., 9.51 million) and the number of unlabeled texts (e.g., 0.969 million). There are two Bengali benchmarks text classification datasets that have been used in this analysis, e.g., BARD [6] and IndicNLP [7]. The two datasets were randomly split into 70:30 ratio, e.g., 70% samples used at training phase and 30% samples used at the testing phase. The datasets statistics summary is shown in Table 3.

The maximum number of category six used at IndicNLP dataset, but the maximum number of training (e.g., 263,299) and testing (e.g., 112,921) sample used in BARD dataset.

4.2 *Evaluation Measures*

The text classification is evaluated in two phases, e.g., embedding model evaluation (e.g., intrinsic evaluation) and test dataset performance evaluation. The embedding models are evaluating with the Spearman ($\hat{\rho}$) and Pearson ($\hat{r}$) correlations [14]. Testing phase evaluation measures such as macro-average precision (M_{ap}), macro-average recall (M_{ar}), weighted average precision (W_{ap}), weighted average recall (W_{ar}), accuracy (Ac), and an confusion matrix (CM) are evaluated the system performance [1].

Table 4 Intrinsic evaluation summary for 100 semantic and syntactic word similarities with the spearman ($\hat{\rho}$) & pearson ($\hat{r}$) correlation

Embedding model	ED	CW	Semantic (%)		Syntactic (%)	
			$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$	$\hat{r}$
AEHT-GloVe	**300**	**11**	**62.78**	**63.10**	**67.19**	**71.97**
MEHT-GloVe	250	15	61.69	62.05	65.30	70.14
AEHT-FastText (Skip-gram)	**300**	**11**	**57.89**	**56.17**	**44.32**	**45.01**
MEHT-FastText (Skip-gram)	250	15	56.16	55.93	43.14	44.05
AEHT-Word2Vec (Skip-gram)	**300**	**11**	**44.90**	**45.10**	**31.97**	**30.32**
MEHT-Word2Vec (Skip-gram)	250	15	43.87	44.01	30.63	30.06

5 Result and Discussion

The embedding model tuning and model performance have been evaluated using intrinsic and extrinsic measures.

Intrinsic Performance: The intrinsic evaluators evaluate the word level performance, whereas the extrinsic evaluators evaluate the sentence or document level performance. Table 4 shows the automatic embedding hyperparameters tuning (AEHT) and manual embedding hyperparameters tuning (MEHT) for 100 semantic as well as syntactic performance. The intrinsic evaluation finds out that the AEHT-GloVe achieved the maximum semantic accuracies of 62.78 and 63.10% for Spearman ($\hat{\rho}$) and Pearson ($\hat{r}$) correlation whereas 67.19 and 71.97% accuracies for Spearman ($\hat{\rho}$) and Pearson ($\hat{r}$) correlation of syntactic dataset. The automatic embedding hyperparameters tuning method achieved better performance compared to manual tuning because the manual tuning does not consider all combinations of ED, CW, and intrinsic evaluation.

Extrinsic Performance: Now, the best performing embedding model is AEHT-GloVe and four classification methods, e.g., CNNs [5], LSTM [12], BiLSTM [11], and SGD [3] used to measure the classification performance (e.g., extrinsic performance). The classification methods hyperparameters have taken from the previous studies. The classification performance summarized in Table 5. The maximum accuracy of 95.16% attained for the AEHT-GloVe+CNNs model from BARD datasets. In contrast, the maximum accuracy of 86.41% was achieved from the IndicNLP dataset using the AETH-GloVe+CNNs model. The minimum accuracies are 81.74% for BRAD and 72.00% for IndicNLP using statistical classification model (e.g., AEHT-GloVe+SGD). The CNNs classifier achieved better performance due to sentence and document level, more correlated feature extracted in this technique, and well-tuned hyperparameters settings.

Table 5 Text classification performance summary for 30% test datasets of BARD & IndicNLP

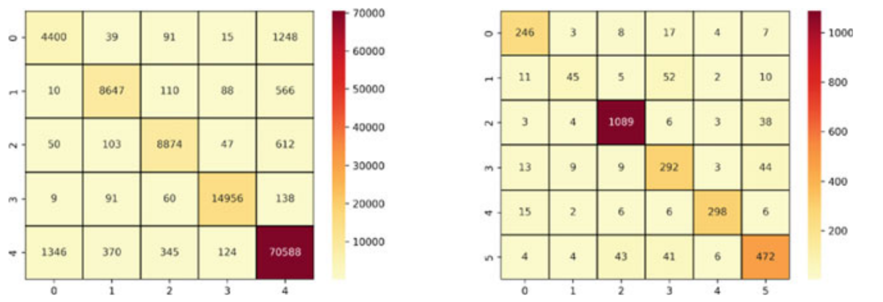
Datasets	Model name	M_{ap} (%)	M_{ar} (%)	W_{ap} (%)	W_{ar} (%)	Ac (%)
BARD	Proposed (AEHT-GloVe+CNNs)	**93.00**	**91.00**	**95.00**	**95.00**	**95.16**
	AEHT-GloVe+LSTM	91.00	91.00	93.00	93.00	93.17
	AEHT-GloVe+BiLSTM	92.00	93.00	94.00	94.00	94.07
	AEHT-GloVe+SGD	78.00	79.00	81.00	81.00	81.74
IndicNLP	Proposed (AEHT-GloVe+CNNs)	**82.00**	**78.00**	**86.00**	**86.00**	**86.41**
	AEHT-GloVe+LSTM	80.00	77.00	84.00	84.00	84.70
	AEHT-GloVe+BiLSTM	81.00	78.00	85.00	85.00	85.93
	AEHT-GloVe+SGD	70.00	69.00	74.00	74.00	74.00

Table 6 Performance comparison summary

Datasets	Model name	Accuracy (%)
BARD	Word2Vec+SGD [3]	79.95
	Word2Vec+BiLSTM [11]	92.05
	Char-embedding+LSTM [12]	91.10
	Proposed (AEHT-GloVe+CNNs)	**95.16**
IndicNLP	Word2Vec+SGD [3]	70.28
	Word2Vec+BiLSTM [11]	82.07
	Char-embedding+LSTM [12]	81.23
	Proposed (AEHT-GloVe+CNNs)	**86.41**

Text Classification Performance Comparison: There are some research which has been conducted on the Bengali text classification system. However, most researches evaluate the model performance using their datasets and have not published their research materials. Few texts classification studies have recently been carried out on BARD and IndicNLP datasets, but they have not used the whole BARD dataset samples and have not split the train and testing set with a standard ratio. So in this study, we take their model's hyperparameters and train and test with the standardized version of BARD and IndicNLP. Table 6 shows the comparison summary.

The Word2Vec+BiLSTM method achieved an accuracy of 92.05 and 82.07% for BARD and IndicNLP datasets which is smaller than the proposed method. This approach is unable to extract word semantic and syntactic features using Word2Vec embedding [11]. The Char-Embedding+LSTM model was not able to extract the



(a) BARD Confusion Matrix (b) IndicNLP Confusion Matrix

Fig. 3 Confusion matrix for AEHT-CNNs model with BARD and IndicNLP dataset. BARD dataset categories with index: economy (0), entertainment (1), international (2), sports (3), and state (4). IndicNLP categories with index: entertainment (0), international (1), Kolkata (2), national (3), sports (4), and state (6)

sentence level better feature concerning CNNs [12]. The statistical model failed to learn the word and document level semantics features [3]. So, based on the previous Bengali text classification research, the AEHT-GloVe+CNNs achieved the best performance based on BARD and IndicNLP datasets.

Error Analysis: Error analysis provides the classification models strength and weaknesses. In this study, the error analysis has conducted using a confusion matrix. Figure 3, shows the confusion matrix of BARD and IndicNLP datasets using the best performing proposed model (e.g., AEHT-GloVe+CNNs). The confusion matrix diagonal entries represent the correct prediction number, and another cell indicates the wrong prediction number. For the BARD datasets, the maximum error rate is 24.04% for the economy category (e.g., 0 index), whereas the minimum error rate is 1.95% for the sports category. The error occurred due to joint data distribution between the category economy (0) and state (4).

For the IndicNLP dataset, the maximum error rate of 64.00% was found for the international (1) category, and the minimum error rate is 5.00% for Kolkata (2) category. Most data distributions in the International (1) category are shared with the National (3) category. The BARD corpus total classification error is 4.84, and the IndicNLP absolute classification error is 13.59. Due to class naming issues, e.g., BARD has two confusing categories. The first one is international, and the second one is state. The other three categories, e.g., economy, entertainment, and sports news, may be overlap with the different three categories (e.g., international and state). For a similar reason, IndicNLP obtained a higher classification error. The shared data distribution is not properly distinguished by the CNNs features extractors. So, the error analysis concludes that the model weakness occurred when the class distribution is shared a standard distribution, whereas the strengths are the model able to carry out the best embedding hyperparameters and extract the semantic and syntactic feature.

6 Conclusion

This paper presented an empirical investigation on optimizing embedding hyperparameters and their impact on Bengali text classification tasks. The results confirmed that the automatic hyperparameters tuning method performed better than the manual optimization. The intrinsic and extrinsic evaluators evaluated the embedding model performance. The intrinsic evaluators evaluated the word level semantic and syntactic feature, whereas the extrinsic evaluators evaluated the sentence or document level features. The proposed method attained the maximum accuracy of 95.16% for the BARD dataset and 86.41% for the IndicNLP dataset. Although the proposed technique showed reasonable outcomes with two hyperparameters (ED, CW), the future study explores the impact of other hyperparameters. Moreover, the current implementation only considered the non-contextual feature models (e.g., GloVe, Word2Vec, & FastText). The multilingual transformer-based language models, including BanglaBERT, m-BERT, and XLM-Roberta, will be investigated in future.

Acknowledgements This work was supported by the Project of Establishment of Sheikh Kamal IT Business Incubator in CUET, BHTPA, ICT Division, Bangladesh.

References

1. Bengali text document categorization based on very deep convolution neural network. Expert Syst Appli **184**, 115394 (2021)
2. Hossain, M.R., Hoque, M.M.: Semantic meaning based Bengali web text categorization using deep convolutional and recurrent neural networks (dcrnns). In: Internet of Things and Connected Technologies, pp. 494–505. Springer International Publishing, Cham (2021)
3. Hossain, M.R., Hoque, M.M.: Automatic Bengali document categorization based on word embedding and statistical learning approaches. In: Proceeding of IC4ME2, pp. 1–6. Rajshahi, Bangladesh (2018)
4. Hossain, M.R., Hoque, M.M., Sarker, I.H.: Text classification using convolution neural networks with fasttext embedding. In: Proceeding of HIS, pp. 103–113. Springer International Publishing, Cham (2021)
5. Hossain, M.R., Hoque, M.M., Dewan, M.A.A., Siddique, N., Islam, N., Sarker, I.H.: Authorship classification in a resource constraint language using convolutional neural networks. IEEE Access **9**, 100319–100338 (2021). https://doi.org/10.1109/ACCESS.2021.3095967
6. Alam, M.T., Islam, M.M.: Bard: Bangla article classification using a new comprehensive dataset. In: 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–5 (2018). https://doi.org/10.1109/ICBSLP.2018.8554382
7. Kunchukuttan, A., Kakwani, D., Golla, S., C., G.N., Bhattacharyya, A., Khapra, M.M., Kumar, P.: Ai4bharat-Indicnlp corpus: monolingual corpora and word embeddings for Indic languages. CoRR abs/2005.00085. https://arxiv.org/abs/2005.00085 (2020)
8. Pennington, J., Socher, R., C.M.: Glove: global vectors for word representation. In: Proceeding of EMNLP, pp. 1532–1543. ACL, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1162
9. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Tran. ACL **5**, 135–146 (2017). https://doi.org/10.1162/tacl_a_00051
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space, pp. 1–12 (2013)

11. Humaira, N., Afia, H., Haque, S.: A study on machine learning and deep learning methods using feature extraction for Bengali news document classification. In: 2021 Asian Conference on Innovation in Technology (ASIANCON), pp. 1–8 (2021). https://doi.org/10.1109/ASIANCON51346.2021.9544761
12. Rahman, M.M., Sadik, R., Biswas, A.A.: Bangla document classification using character level deep learning. In: 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–6 (2020). https://doi.org/10.1109/ISMSIT50672.2020.9254416
13. Witt, N., Seifert, C.: Understanding the influence of hyperparameters on text embeddings for text classification tasks. In: Research and Advanced Technology for Digital Libraries, pp. 193–204. Springer International Publishing, Cham (2017)
14. Hossain, M.R., Hoque, M.M.: Towards Bengali word embedding: corpus creation, intrinsic and extrinsic evaluations. In: Proceedings of the 17th International Conference on Natural Language Processing (ICON), pp. 453–459. NLP Association of India (NLPAI), Indian Institute of Technology Patna, Patna, India (2020)

Healthcare

Design of Novel Feature Union for Prediction of Liver Disease Patients: A Machine Learning Approach

Rubia Yasmin, Ruhul Amin, and Md. Shamim Reza

Abstract Feature engineering is treated as a crucial step in any intelligent system pipeline and practitioners often spend 70–80% of their time in this phase before modeling. Especially, in machine learning (ML), the performance of the classification systems deteriorates as irrelevant features are added, even when the features presented contain enough information about the problem. ML researchers often use feature selection and feature extraction techniques individually to filter out the redundant information from the data, but would not get desired results or performance due to choosing a single method only. In this article, we take an attempt to combine both feature selection and feature extraction methods that give an improvement in the prediction performance than using a single feature selection or feature extraction method. In the application of the proposed method, we consider the Indian liver disease patient dataset to classify liver disease using machine learning classifiers. In addition, ensemble classifiers also take into account reproducing and robustifying classification accuracy. The performance is evaluated and compared with different metrics such as accuracy, precision, recall, and F1-score. Using the proposed method, the KNN classifier can diagnosis 76.03% with a precision of 76.69%, recall 96.23%, and 85.36% F1-score which is much better than the individual feature selection and feature extraction approach. The proposed method can also be extended to medical datasets other than liver disease so that these types of deadly diseases can be correctly diagnosed for the betterment of humanity.

Keywords Machine learning · Feature extraction · Feature selection · Feature union · Liver disease

R. Yasmin (✉) · R. Amin · Md. S. Reza
Department of Statistics, Pabna University of Science & Technology, 6600 Pabna, Bangladesh
e-mail: rubiayasmin0179@yahoo.com

R. Amin
e-mail: ruhulstat6@gmail.com

Md. S. Reza
e-mail: shamim.reza@pust.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_36

1 Introduction

The information from different disease patients is massive and complex. An early and accurate diagnosis of a disease is very important in medical science. We can use this disease information in prediction, and this also helps in the treatment plan. Different machine learning techniques can be used to predict the disease by using the feature information, then time and cost are saved. A massive amount of data affects the performance of machine learning classifiers. Feature selection and feature extraction are a dimension reduction technique that aims at filtering out the redundant information from the data. Different techniques select their features differently because there has no exact procedure to select the most informative variables. The different researcher considers feature selection based on one chosen method only [1]. This different chosen feature selection procedure produces different results. As a result, we use a combination of feature selection and feature extraction strategies to increase prediction performance.

The liver is the largest solid organ of the body which is shaped like a cone. A person cannot survive without a functioning liver. If any disturbances occur in liver function, then it causes illness that is known as liver disease which is also called hepatic disease. The liver performs more than 500 functions, and some well-known functions are the production of bile, production of important proteins for blood clotting, purification of blood, helping in fat digestion, decomposing red blood cells, and detoxifying harmful chemicals [2]. Liver disease is a broad term that covers all the potential problems that cause the liver fails to perform its designated function. Usually, more than 75% or three-quarters of liver tissue needs to be affected before a decrease in function occurs [3]. Liver cancer, liver cirrhosis, hepatitis A, hepatitis B, hepatitis C, and liver failure are common liver diseases. Liver disease cannot be identified at the initial stage because it grows without any symptoms.

The main objectives of this study are to predict liver disease accurately and reduce the time and cost of diagnosis of the disease. To classify the liver patient accurately, we use six different machine learning classifiers as logistic regression, K-nearest neighbor, support vector machine, random forest, multilayer perceptron, ensemble learning, and their performance is compared with different measures as accuracy, precision, recall, and F1-score. Receiver operating characteristic (ROC) curve is also used to compare the performance of classifiers.

2 Related Work

Singh et al. [1] have used the classification models such as logistic regression, Naïve Bayes, SMO, IBK, J48, random forest in their research work. They used the ILPD dataset from UCI repository. After applying the feature selection method, the logistic regression classifier gives an accuracy of 74.36% which is best. In 2014, Gulia et al. [4] have used the Indian liver patient dataset to classify the disease patients. This research shows SVM which gives better results with an accuracy of 71.36% before applying feature selection methods. After applying the feature selection method, random forest algorithms give the best result with an accuracy of 71.87%. Razali et al. [5] in their research paper used two regression algorithms and two classification algorithms. In their work, the linear regression algorithm in liver disease dataset gives better results than the Poisson regression algorithm. Neural networks and Bayes point machines are used as classification algorithm. The overall best result gives the Bayes point machines for solving the problem relating to liver disease. It gives an accuracy of 70.52% which is the best accuracy of other algorithms.

Ramana et al. [6] in their research work obtained 51.59% accuracy on the Naïve Bayes classifier, 66.66% on BPNN, 62.60% on KNN, 55.94% on C4.5, and 62.60% on SVM. They also compute other performance measures as precision, sensitivity, and specificity. Recently, Azam et al. [7] worked on the ILPD dataset. The authors applied five different classifiers to the dataset with or without applying the feature selection technique. SVM gives the best result (71.22%) without applying the feature selection technique, and K-nearest neighbors give the best result than all other classifiers with an accuracy of 74.15% after applying feature selection.

3 Dataset Description

Indian liver patient dataset (ILPD) is used here to diagnose liver patients which are collected from the University of California Irvine (UCI) machine learning repository [8]. This dataset was originally collected from the northeast of Andhra Pradesh, India. This dataset consists of 583 patients, whereas 75.64% are male and 24.36% female. There have 416 disease cases and 167 non-disease cases shows in Fig. 1. ILPD dataset contains 11 features, whereas one is considered an output feature. The features are Age of the patients, Gender of the patients, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, Albumin and Globulin Ratio and Target which is the selector field used to split data into two sets. Only the Gender feature is categorical. There have 4 missing values in Albumin and Globulin Ratio feature. The target variable 1 indicates disease, and 2 indicates non-disease.

4 Feature Reduction and Classification Techniques

Feature selection and feature extraction are two major approaches for dimension reduction. In feature selection, the subset of features is chosen from original features, whereas feature extraction compromises the projection of original features in lower dimensions.

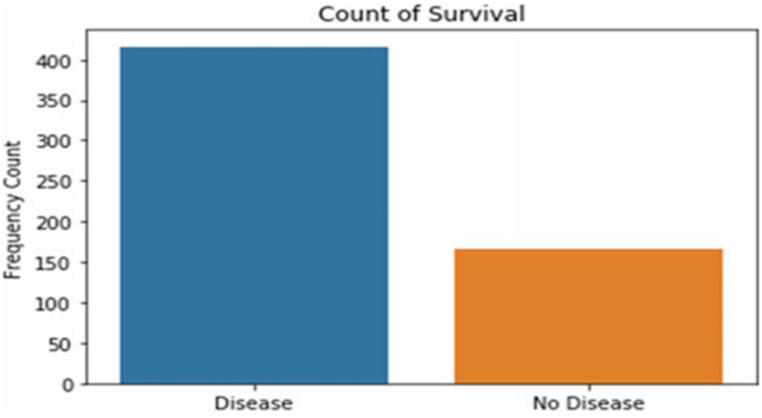


Fig. 1 Count plot of the target variable

4.1 Feature Selection Techniques

Feature selection is an important task in the field of machine learning as well as in data science. It is a way to select the most informative features among many features. The optimal feature selection removes the complexity of the system and increases the reliability, stability, and also classification accuracy [9]. Statistical-based feature selection methods evaluate the relationship between each input variable and target variable using statistical techniques. It also selects those input variables which have the strongest relationship with the target variable [10]. We use the mutual information feature selection technique which is described below.

- **Mutual Information Feature Selection**

In the field of information theory, mutual information is used to select the most informative features of the given datasets. Mutual information is a measure between two random variables X and Y that quantifies the amount of information obtained about one random variable, through the other random variable. The mutual information is given by

$$\mathrm{MI}(x, y) = \int_x \int_y P(x, y) \log \frac{p(x, y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y$$

where p(x, y) indicates the joint probability density function of X and Y. p(x) and p(y) are the marginal density function. The mutual information determines how similar the joint distribution p(x, y) is to the products of the factored marginal distributions [11].

4.2 Feature Extraction Techniques

Feature extraction aims to reduce dimensions by building new values (features) from the original features of the dataset. The created features are expected to contain summarized information of the original feature sets. Therefore, a complex dataset can be represented by a reduced set of features, which yet describe redundant information of the initial set of features. Kernel principal component analysis technique is used here as a feature extraction method.

- **Kernel PCA**

Kernel PCA can be considered as a nonlinear form of normal PCA. We run the data through a kernel function and project it into a new high-dimensional feature space momentarily. Now, we separate the classes by a straight line. Then, the algorithm projects the data into a lower-dimensional space by using the normal PCA. Different types of kernels such as 'linear', 'rbf', 'poly', and 'sigmoid' are used in kernel PCA [12].

Recently, the feature union technique combines several feature selection and extraction mechanisms into a single transformation. Usually, it produces a larger dataset on which a model can be trained. For fitting feature union, each feature selection and extraction mechanism are applied to the data independently. These transformers are applied in parallel. There has no way of checking whether the two transformers might produce identical features. The feature sets are disjoint in the feature union method [13].

4.3 Classification Algorithm

Different classification algorithms have been studied in the machine learning community. Most of the common learning algorithms are discussed below:

- **Logistic Regression**

Logistic regression (LR) is a classification algorithm that is a linear model, used to predict a binary output, i.e., yes/no, pass/fail, 0/1, and so on. LR algorithms are more robust and also applicable for nonlinear datasets which is the main advantage [14]. Let, there are N input features like X_1, X_2, … $X_{N,}$ and P is the probability of occurring the event, and 1-P indicates the probability of not occurring the event. Then, the mathematical expression of N features of the model is in the form:

$$\text{Logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N \quad (1)$$

where β_0 indicates the intercept term, and β_i ($i = 1, 2, \ldots, N$) is the regression coefficient.

- **K-Nearest Neighbor (KNN)**

KNN is one of the most fundamental distance-based algorithms. The KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the *k* closest entries in the training set are considered further. The new item is then placed in the class that contains the next items from this set of *k* closest items.

- **Support Vector Machine (SVM)**

Support vector machines are supervised learning models with incorporated learning algorithms that analyze data and deduce patterns, used for classification and regression [5]. A support vector machine constructs an optimal hyperplane or set of hyperplanes in a high or infinite-dimensional space that has the largest distance to the nearest training data point of any class can achieve good separation. The empirical classification errors are minimized, and the margin is maximized in SVM. The classification SVM type-1 model minimizes the error function [15]:

$$\frac{1}{2}w^T w + c\sum_{i=1}^{N}\zeta i \tag{2}$$

And the classification SVM type-2 model minimizes the error function below:

$$\frac{1}{2}w^T w - v_p + \frac{1}{N}\sum_{i=1}^{N}\zeta i \tag{3}$$

- **Random Forest (RF)**

Random forest classifier is an ensemble supervised machine learning method. It predicts the target variable by combining many decision trees usually trained with the 'bagging' method. This classifier may smoothly handle big data with missing values and be used to predict both the regression and classification problems. The training algorithm for random forests applies the general technique of bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For, $b = 1, 2, \ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} fb(x')$$

or by taking the majority vote in the case of classification trees [16].

- **Multilayer Perceptron (MLP)**

The multilayer perceptron (MLP) is one of the most common neural network models in the fields of deep learning. A multilayer perceptron (MLP) is a class of feedforward artificial neural networks (ANN). MLP consists of three layers of nodes: an input layer, a hidden layer, and an output layer. MLP is a relatively simple form of a neural network because the information travels in one direction only. Each node uses a nonlinear activation function except for the input nodes**.** A linear activation function maps the weighted input to the output of each neuron. The two common activation functions are both sigmoid and are described by

$$y(v_i) = \tanh(v_i) \text{ and } y(v_i) = \left(1 + e^{-vi}\right)^{-1}$$

Here, y_i is the output of the i-th node (neuron), and v_i is the weighted sum of the input connections [17].

- **Ensemble Learning Classifier**

In machine learning, the algorithm ensemble model combines several base models that give one optimal predictive model which have better performance. There have three dominant classes of ensemble learning methods as bagging, stacking, and boosting. Bagging or bootstrap aggregation is the most powerful and simple effective method which generally considers many decision trees on different samples of the input dataset and then averaging the results of predictors. Stacking is an advanced ensemble learning technique that uses different learning models on the same dataset to build a new model. A set of base-level classifiers is generated in the first phase and in the second phase, a meta-level classifier is learned which combines the base-level predictions. Boosting methods add different ensemble members sequentially and correct the prediction errors then use the simple voting or averaging method to combine the predictions [18].

5 Methodology

In methodology, before applying our proposed feature reduction method, we extensively investigate different data pre-processing techniques. Data pre-processing is required to improve the data quality and prepare the raw data to make it suitable for machine learning algorithms [19]. Most of the real-world datasets need to be pre-processed. Every dataset is different and creates unique challenges. In our ILPD dataset, only 'gender' is a categorical feature. We encode it by the ordinal encoder Scikit-learn package. The only 4 missing values in Albumin and Globulin Ratio are replaced by the median. Boxplots are used to identify the outliers in our dataset. The features such as Alkaline Phosphatage, Alamine Aminotransferase, and Aspartate

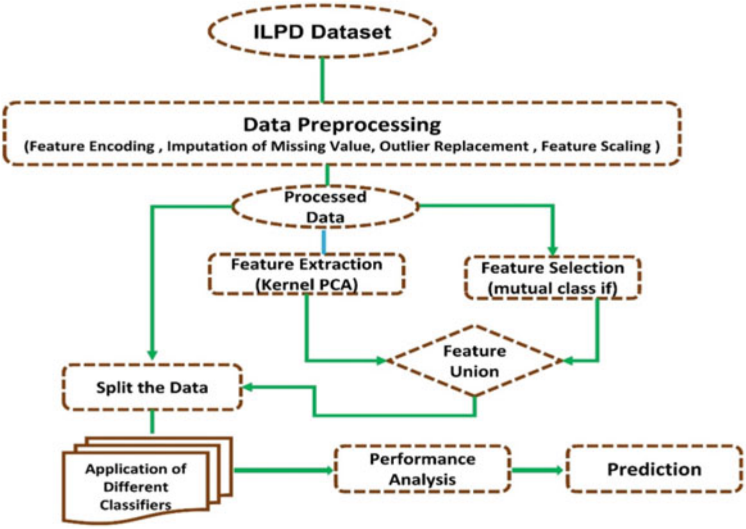


Fig. 2 Proposed feature reduction strategy

Aminotransferase contain a greater number of outliers. After identifying outliers, it fills in by mean. We check the outliers by z-score value. If an outlier exists, it has been replaced by the median of this feature. StandaredScaler is used for scaling the features. We use the methodology followed in Fig. 2, for conducting liver patient dataset prediction experiments. Redundancy of the data is removed by using feature selection and feature extraction techniques. The feature selection process based on mutual feature information and the feature extraction process based on the kernel PCA are considered here for selecting the redundant features in the dataset. We use the feature union pipeline for our proposed combination method. This proposed feature combination helps us to find out the most redundant and representative features that are selected from both feature selection and extraction methods.

6 Evaluation Protocols, Result, and Discussion

The performance matrices such as accuracy, precision, recall, and F-score are used in comparing the results of different classifiers. These measures are calculated from the confusion matrix. The confusion matrix summarizes the actual and predicted results of a classification model [20]. A confusion matrix is shown in Table 1.

Where TP is the number of correctly classified positive instances, FP is the number of misclassified positive instances, FN is the number of misclassified negative instances, TN is the number of correctly classified negative instances. We can find different performance measures from the given relationship:

Table 1 Evaluation matrices

Data class	Predicted true	Predicted false
Actual true	TP	FN
Actual false	FN	TN

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 Score} = \frac{2 * \text{TP}}{2 * (\text{TP} + \text{FP} + \text{FN})}$$

This paper represents how different feature selection methods and feature extraction methods are combinedly used to accurately classify liver patients into two classes as disease and non-disease. We use the feature union pipeline to combine both effects of feature selection and extraction. Kernel PCA and mutual feature information are used here as feature extraction and feature selection techniques, respectively.

The performance of different classification algorithms is compared before and after applying feature union techniques. Six machine learning classifier is used to classify the liver patient dataset. Our study shows that if we replaced the missing value and outliers by median and select the most informative features by mutual information feature selection method and add the effects of Kernel PCA, then classification performance is improved.

We split the dataset into the train (75%) and test (25%) set after processing the dataset. According to Table 2, the K-neighbor classifier gives an accuracy of 74.66%, which is the best among the other classifiers without applying a feature union pipeline. After using the mutual information feature selection technique, we select the 5 most important features for training the dataset. Then, we select 5 principal components for kernel PCA. After applying our proposed feature union, overall, we get better performance of all the classifiers except the random forest classifier. From Table 3, the KNN classifier gives the best accuracy of 76.03%. So, we can say the KNN classifier outperformed all other classifiers after selecting the most important feature from the mutual information feature selection method and combining the effects of kernel PCA via feature union pipeline.

ROC curve is a plot that summarizes the performance of a binary classification model on the positive class. It is a two-dimensional graph, where the FP rate in the X-axis and TP rate in the Y-axis. Figure 3 shows that almost all of the classifier covered more area than other classifier using the proposed technique. The KNN algorithm provides 76.03% accuracy, which is better than the other classifier using the suggested technique, according to result comparison in Fig. 4.

Table 2 Comparing table of the results of different classifiers on ILPD without feature union

Classifier	Accuracy	Precision	Recall	F1-score
Logistic regression	67.12	71.01	92.45	80.33
KNN	74.66	77.60	91.51	83.98
SVM	72.60	72.60	99.99	84.13
Random forest	73.29	75.57	93.40	83.54
MLP	62.33	73.83	74.53	74.18
Ensemble	72.60	73.91	96.23	83.61

Table 3 Comparing table of the results of different classifiers on ILPD after applying selected feature union pipeline

Classifier	Accuracy	Precision	Recall	F1-score
Logistic regression	71.92	72.73	83.93	83.53
KNN	76.03	76.69	96.23	85.36
SVM	72.60	72.60	99.99	84.13
Random forest	71.23	74.62	91.51	82.20
MLP	66.44	78.79	73.58	76.10
Ensemble	69.86	73.13	92.45	81.67

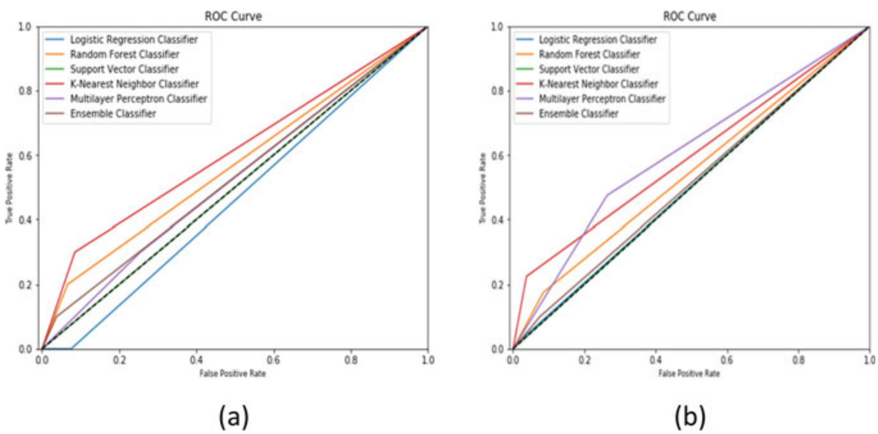


Fig. 3 ROC curve of the dataset (**a**) without feature union, (**b**) with feature union

7 Conclusion

With data collection from feature extraction techniques in classical machine learning systems as well as automatically learned features obtained from deep learning model are often showed high dimensionality. Ingesting these features directly to build

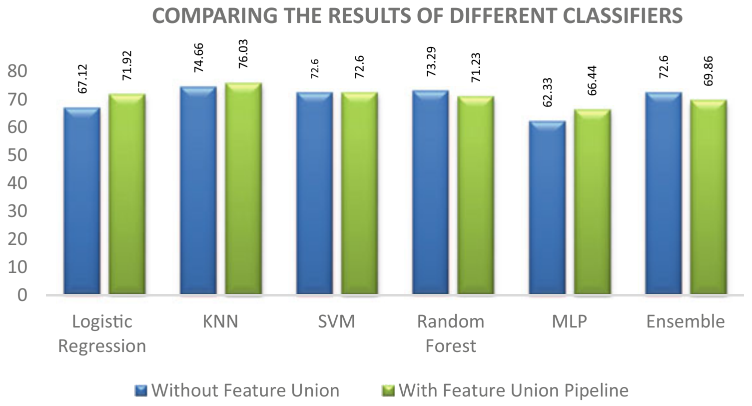


Fig. 4 Comparison chart for different classifier

models would be foolhardy and generally requires a large amount of memory, computation power. In this study, we deal with the potential problem by integrating mutual information feature selection and kernel principal component analysis (KPCA) feature extraction techniques. To test the effectiveness of our methods, we have considered liver patients' data. Based on the accuracy and different evaluation matrices, we may conclude that our method reduces the feature's dimensionality and improves the classification performances considerably. In the application of the proposed system to medical data, it can be used as an adjunct tool to cross-check the diagnosis of liver cancer with the doctor's assessment.

References

1. Singh, J., Bagga, S., Kaur, R.: Software-based prediction of liver disease with feature selection and classification techniques. Procedia Comput. Sci. **1**(167), 1970–1980 (2020)
2. Liver: anatomy and functions | Johns Hopkins Medicine [Internet]. [cited 28 Sep 2021]. Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/liver-anatomy-and-functions
3. Liver disease symptoms, treatment, stages, signs, types, diet [Internet]. [cited 28 Sep 2021]. Available from: https://www.medicinenet.com/liver_disease/article.htm
4. Gulia, A., Vohra, R., Praveen, R.: Liver patient classification using intelligent techniques. Int. J. Adv. Res. Comput. Sci. Softw. Eng. [Internet]. **4**(2), 5110–5115 (2014). Available from: http://archive.ics.uci.edu/ml/
5. Razali, N., Mostafa, S.A., Mustapha, A., Wahab, M.H.A., Ibrahim, N.A.: Risk factors of cervical cancer using classification in data mining. J. Phys. Conf. Ser. **1529**(2) (2020)
6. Venkata Ramana, B., Babu, M.S.P., Venkateswarlu, N.: A critical study of selected classification algorithms for liver disease diagnosis. Int. J. Database Manag. Syst. **3**(2), 101–114 (2011)
7. Azam, S., Rahman, A., Iqbal, S.M.H.S., Ahmed, T.: Prediction of liver diseases by using few machine learning-based approaches. Aust. J. Eng. Innov. Technol. **2**(5), 85–90 (2020)

8. UCI machine learning repository: ILPD (Indian Liver Patient Dataset) Data Set [Internet]. [cited 28 Sep 2021]. Available from: https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29
9. Why, How and When to Scale your Features | by Sudharsan Asaithambi | GreyAtom | Medium [Internet]. [cited 28 Sep 2021]. Available from: https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e
10. How to choose a feature selection method for machine learning [Internet]. [cited 28 Sep 2021]. Available from: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
11. Mutual information-based feature selection | Thomas Huijskens [Internet]. [cited 28 Sep 2021]. Available from: https://thuijskens.github.io/2017/10/07/feature-selection/
12. 11 Dimensionality reduction techniques you should know in 2021 | by Rukshan Pramoditha | Towards Data Science [Internet]. [cited 28 Sep 2021]. Available from: https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b
13. sklearn.pipeline.Pipeline—scikit-learn 1.0 documentation [Internet]. [cited 28 Sep 2021]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html
14. Maniruzzaman, M., Rahman, M.J., Al-MehediHasan, M., Suri, H.S., Abedin, M.M., El-Baz, A., Suri, J.S.: Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J. Med. Syst. **42**(5), 1–17 (2018)
15. Rahman, A.K.M.S., Javed Mehedi Shamrat, F.M., Tasnim, Z., Roy, J., Hossain, S.A.: A comparative study on liver disease prediction using supervised machine learning algorithms. Int. J. Sci. Technol. Res. **8**(11), 419–422 (2019)
16. Random forest—wikipedia [Internet]. [cited 28 Sep 2021]. Available from: https://en.wikipedia.org/wiki/Random_forest
17. Multilayer perceptron—wikipedia [Internet]. [cited 28 Sep 2021]. Available from: https://en.wikipedia.org/wiki/Multilayer_perceptron
18. A gentle introduction to ensemble learning algorithms [Internet]. [cited 7 Oct 2021]. Available from: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/
19. Data preprocessing in machine learning: 7 easy steps to follow | upGrad blog [Internet]. [cited 28 Sep 2021]. Available from: https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/
20. Das, R., Sengur, A.: Evaluation of ensemble methods for diagnosing of valvular heart disease. Expert Syst. Appl. [Internet]. **37**(7), 5110–5115 (2010). Available from: https://doi.org/10.1016/j.eswa.2009.12.085

Classifying Humerus Fracture Using X-Ray Images

Tahmina Akter Sumi, Nanziba Basnin, Md. Shahadat Hossain, Karl Andersson, and Md. Sazzad Hoassain

Abstract Bone is the most important part of our body which holds the whole structure of human body. The long bone situated in the upper arm of human body between the shoulder and elbow junction is known as "Humerus". Humerus works as a structural support of the muscles and arms in the upper body which helps in the movement of the hand and elbow. Therefore, any fracture in humerus disrupts our daily lives. The manual fracture detection process where the doctors detect the fracture by analyzing X-ray images is quite time consuming and also error prone. Therefore, we have introduced an automated system to diagnose humerus fracture in an efficient way. In this study, we have focused on deep learning algorithm for fracture detection. In this purpose at first, 1266 X-ray images of humerus bone including fractured and non-fractured have been collected from a publicly available dataset called "MURA". As a deep learning model has been used here, data augmentation has been applied to increase the dataset for reducing over-fitting problem. Finally, all the images are passed through CNN model to train the images and classify the fractured and non-fractured bone. Moreover, different pretrained model has also been applied in our dataset to find out the best accuracy. After implementation, it is observed that our model shows the best accuracy which is 80% training accuracy and 78% testing accuracy comparing with other models.

Keywords Humerus fracture · Deep learning algorithm · Data augmentation · Convolution neural network

T. A. Sumi · Md. S. Hossain (✉)
University of Chittagong, Chittagong, Bangladesh
e-mail: hossain_ms@cu.ac.bd

N. Basnin
International Islamic University Chittagong, Sonaichhari, Bangladesh

K. Andersson
Lulea University of Technology, Skelleftea, Sweden

Md. S. Hoassain
University of Liberal Arts Bangladesh, Dhaka 1209, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_37

1 Introduction

A humerus fracture is a break in the upper arm's humerus bone. Plain radiography is used to identify humerus fractures [22]. To determine the type of fracture, its anatomical location followed by fragmentation and displacement levels is taken into account. Albeit experienced surgeons may misdiagnose because humerus fracture has various representations [5]. As a result, an effective, as well as accurate method, is required to classify the fracture. Deep learning is a subset of artificial intelligence that extracts features and creates transformations using a cascade of many layers of nonlinear processing units. It is based on the learning of several levels of features or representations of the data. With the success of employing a deep learning model to identify and categorize images, there has been interest in using deep learning in medical image analysis in a variety of domains, including skin cancer diagnosis, diabetic retinopathy, mammographic lesions, and lung nodules identification. Trials in orthopedic surgery and traumatology, on the other hand, are extremely rare, despite their relevance to public health. Therefore, for this research, a CNN model is employed to classify humerus fractures as it proved to perform better in case of image classification [4, 9, 17, 23]. A dataset is collected and preprocessed. Data augmentation is applied to increase the robustness of the dataset. A CNN model is applied to train and test the dataset.

The following sections of this paper are organized as follows: Sect. 2 describes the previous works based on ALL detection. Section 3 depicts the entire methodology where whole CNN model used in this paper has been illustrated broadly. Section 4 is about the the result of our research, and finally, Sect. 5 is about conclusion and future work.

2 Literature Review

This section basically covers the previous research works related to the detection of bone fracture. Sasidhar et al. [20] in 2021 applied 3 pretrained model (VGG16, Densenet121 and Densenet169) on humerus bone images collected from the MURA dataset for detecting the fractured bones . They showed their best accuracy 80% using DenseNet161. Demir et al. [7] introduced an exemplar pyramid feature extraction method for classification of humerus fracture. They employed HOG and LBP for feature generation. They achieved a outstanding accuracy which is 99.12%. However, in this research, only 115 images images have been used which is very poor. Chung et al. [5] in 2018 developed a deep convolution neural network for classifying the fracture types where they gained a promising result with 65–86% top-1 accuracy. Negrillo-Cárdenas et al. [14] in 2020 proposed a geometrically-based algorithm for detecting landmarks in the humerus for the purpose of reducing supracondylar fractures of the humerus. In this research, they calculated the distance between two corresponding landmarks which is 1.45 mm ($p < 0.01$). Sezer and Sezer [21] shoulder images had been evaluated by CNN for the feature extraction and the classification of the head of humerus into three categories, for example, normal, edematous and Hill-Sachs lesion with 98.43% accuracy. Negrillo-Cárdenas et al. [15] in 2019 considered a geometrical approach and spatial for detecting the landmarks on distal humerus. In this paper, they calculated 6 points for each bone. All these research took humerus as parameter.

Deviating from humerus fracture, some researchers also used other bones as parameter, for example, shoulder, femur, calcaneus, etc. De Vries et al. [6] in 2021 worked on predicting the chance of osteoporotic fracture (MOF) where they developed three machine learning models (Cox regression, RSF, and ANN) to show the comparison. Here, they showed that Cox regression outperformed the other models with 0.697 concordance-index. Pranata et al. [16] applied ResNet and VGG for the detection of calcaneus fractures using CT images where they gained a good accuracy 98%. Adams et al. [1] worked on the detection of neck of femur fractures where they implemented deep convolutional neural networks. In this work, the researchers showed an accuracy of 90.5% and AUC of 98%. Devana et al. [8] utilized machine learning model for the prediction of complications following reverse total shoulder arthroplasty (rTSA). Here, they applied five classifiers where XGBoost gained the top AUROC and AUPRC of 0.681 and 0.129.

After analyzing all the previous research work, it can be concluded that, in [7, 20], research on humerus fracture detection has been done, but further improvement is needed here. In [5, 21], humerus fracture classification has been done where they classified the fracture types rather detection. Negrillo-Cárdenas et al. [14, 15] studies were about the detection of landmark on humerus. So, it is clear that there has been very little research on humerus fracture detection. So in this study, we have tried to make a contribution on the detection of humerus fracture using deep learning model (Table 1).

3 Methodology

This section is about the whole network architecture that has been utilized in this study to detect humerus fracture. Figure 1 illustrates the system flowchart of this study.

As we can see in Fig. 1, this study starts with collecting dataset. After the data collection, all the images are sent to the data augmentation section. After that, preprocessing step is applied to the augmented data where the images are resized and then converted to an array. After the completion of pre-processing step, the processed images are sent to the CNN model where the features are extracted, and classification step finally predicts the images as “positive” or “negative”.

3.1 Dataset

The dataset we have used in this research for humerus fracture detection is collected from a public source which is named “musculoskeletal radiographs (MURA)”

Table 1 Related works

Reference	Year	Parameters	Methodology	Performance
[20]	2021	Humerus bone fracture	VGG16, Densenet121, Densenet169	Accuracy: 0.80
[7]	2020	Humerus fractures	HOG, LBP	Accuracy: 99.12
[5]	2018	Proximal humerus fractures	DCNN	Accuracy: 0.65–0.86
[14]	2020	Landmarks in the humerus	Geometrically-based algorithm	Distance between two corresponding landmarks 1.45 mm ($p > 0.01$)
[21]	2020	Diagnosis of proximal humerus	CNN	98.43
[15]	2019	Distal humerus	Geometrical approach	Calculated 6 points for each bone
[6]	2021	Hip, wrist, spine, and humerus fractures	Cox regression, RSF and ANN	Concordance-index 0.697 for Cox regression
[16]	2019	Calcaneus	ResNet and VGG	Accuracy: 0.98
[1]	2019	Neck of femur	Deep convolutional neural networks	0.91 (accuracy) 0.98 (AUC)
[8]	2021	Shoulder arthroplasty	LR, XGBoost, RF	AUROC:681, AUPRC: 0.129 (for XGBoost)

dataset. The MURA dataset is a large dataset containing 40,561 images collected from 12,173 patients which is labeled by board-certified radiologists of the Stanford Hospital whether it is positive or negative. To inquire the different types of abnormalities present in the dataset, the report has been reviewed manually hundred abnormalities where fifty three studies have been marked as fractures, 48 has been marked with hardware, 35 has been marked with degenerative joint diseases, and 29 has been marked with other mixed abnormalities.

Anyway, this dataset contains 7 types of bone fracture images, namely elbow, finger, forearm, hand, humerus, shoulder, and wrist. As we are working with the detection of humerus fracture, so only the X-ray images of humerus fracture have been considered in this study. Total of 1266 (669 negatives and 597 positives) X-ray images of humerus have been used in MURA dataset which are collected from 727 patients.

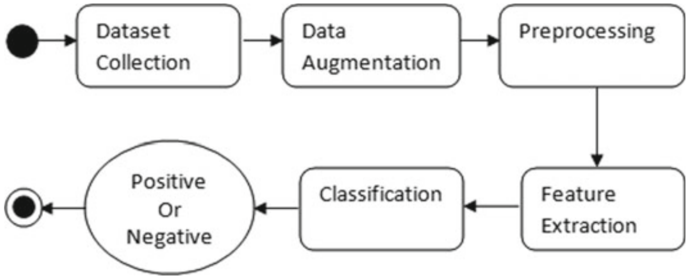


Fig. 1 System flowchart

3.2 Data Augmentation

Enhancing the performance of a model is an all time challenging task. Data augmentation helps to a great extent in this case. Generally by the data augmentation, the dataset is increased by creating altered copies of data which are already exist. It is about fabricating more data from the existing dataset without losing the data information. Data augmentation lessens the risk of over-fitting problem by increasing dataset. However, data augmentation can boost the model accuracy when the dataset is insufficient.

As stated earlier, in this study, we have used total of 1266 X-ray images of humerus including positive case and negative case which is relatively smaller for deep learning method. That is why we have implemented the data augmentation process on our dataset before training the model. For this augmentation, ImageDataGenerator has been utilized here. The settings for image augmentation that we utilized have been illustrated in Table 2.

Table 2 Data augmentation settings

Augmentation techniques	Range
Rotation	25
Width shift	0.10
Height shift	0.10
Zoom	0.40
Horizontal flip	True
Fill mode	Nearest

3.3 Pre-processing

Image pre-processing creates a positive effects on image by increasing the quality of feature extraction. Image pre-processing is needed to clean the image dataset prior to train the model and inference. It generally works by removing background noise, reflection of light, normalizing the images, and preparing the images for better feature extraction.

In this study, the pre-processing step starts by resizing the images into 96×96 dimension. After that, the pixel values of the images are converted to an array form where each of the element of array indicates the value of pixel. The pixel value range is in between 0 and 255 which creates a great variation among the array elements. To reduce this difference of this array elements, each of the pixel values are divided by 255 so that all the array elements come to a range between 0 and 1.

3.4 Convolutional Neural Network Model

Convolutional neural network model (CNN) and a deep learning neural network model have been utilized in this study for training the image dataset as CNNs are effective in learning the raw data automatically [2, 3, 19, 26]. The architecture of this model has been illustrated in Fig. 2.

According to Fig. 2, the model consists of three convolution layers where there are 32 filters in first convolutional layer, 64 filters in second convolutional layer, and 128 filters filters in third convolutional layer with 3×3 kernel size. In Eq. (1), the mathematical expression of convolution layer over an image $s(x, y)$ is defined using filter $t(x, y)$.

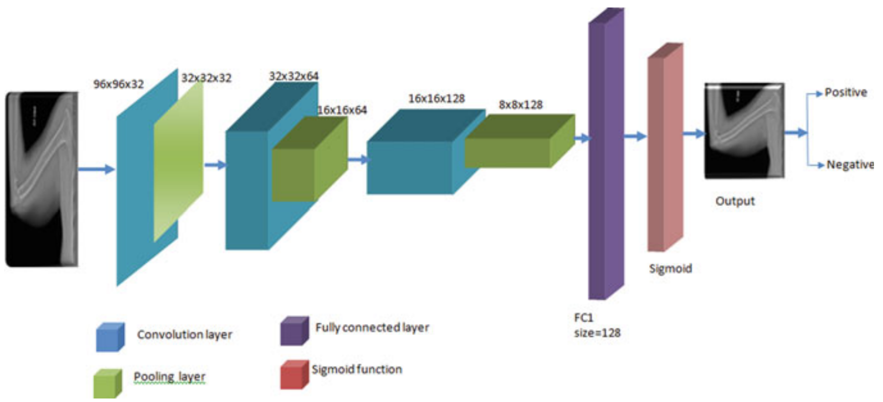


Fig. 2 Model architecture

$$t(x, y)s(x, y) = \sum_{y=-j}^{j} \sum_{y=-k}^{k} w(m, n) f(x - m, y - n) \tag{1}$$

Each of the convolution layer is followed by batch normalization and ReLU activation function. The ReLU activation function is expressed mathematically which is shown in Eq. (2)

$$\mathrm{Relu}(z) = \max(0, z) \tag{2}$$

A max pooling layer with 2 × 2 pool size has been included in each convolution layer. The max pooling layer efficiently downsizes the tight scale images and mitigates total numbers of parameters. Thus, it reduces the size of the dataset.

The architecture is the followed by fully connected (FC) layer where all the inputs of 1 layer are linked to each of the activation unit of the subsequent layer. A dropout layer is attached to the FC layer, and here, the size of the FC layer is 128. The dropout layer helps to reduce the over-fitting problem by randomly disconnecting nodes. Here, we have selected 0.2 as dropout range for this layer.

Finally, a sigmoid function has been applied after the FC layer which acts as classifier to classify images as "positive" or "negative". The mathematical expression of sigmoid function has been shown in the following Eq. (3).

$$\mathrm{Sigmoid}(x) = \frac{1}{1 + \mathrm{e}^{-\theta^T x}} \tag{3}$$

4 Experimental Results and Evaluation

The outcomes after implementing our model on the humerus bone images collected from MURA dataset have been explained in detail in this section. The result after applying five fold cross validation has been illustrated. In addition, we have also shown the comparison of our proposed model with other traditional models.

4.1 Results and Discussion

Figure 3 graphically represents the learning curve of our model. According to Fig. 3, we can see that the X axis denotes the total epoch used to train dataset, whereas the Y axis denotes the accuracy and loss, respectively. From Fig. 3a, it can be observed that the training accuracy increases from 52 to 71%, and testing accuracy increases from 53 to 72%, respectively, after the completion of 50 epochs. Finally, the training accuracy reaches to 82%, and testing accuracy reaches to 80% after the completion of final epoch.

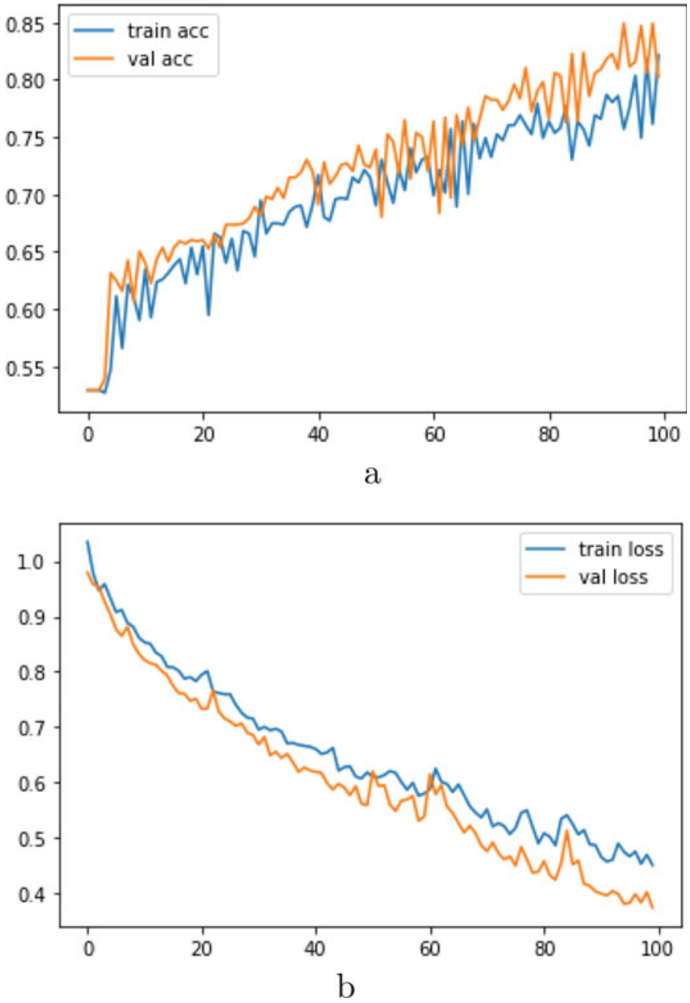


Fig. 3 Accuracy and loss curve: **a** accuracy versus epoch **b** loss versus epoch

Accordingly, from Fig. 3b, it can be observed that the training loss decreases from 1.03 to 0.62, and testing loss decreases from 0.97 to 0.56, respectively, after the completion of 50 epochs. Finally, the training loss reaches to 45%, and testing loss reaches to 37% after the completion of final epoch.

In this research, we have attempted five fold cross validation for assessing our model. In Table 3, the training and testing accuracies after applying five fold cross validation have been demonstrated where individual accuracy for each fold is added. The average (77.20% training and 73.60% testing) and best (80% training and 78% testing) accuracies are also added in the table which are calculated from the five fold cross validation.

Figure 4 illustrates accuracy versus epoch curve for each fold.

Table 3 Data augmentation settings

Fold	Training accuracy (%)	Testing accuracy (%)
1	80	78
2	77	75
3	78	71
4	73	70
5	78	74
Average accuracy	77.20	73.60
Best accuracy	80	78

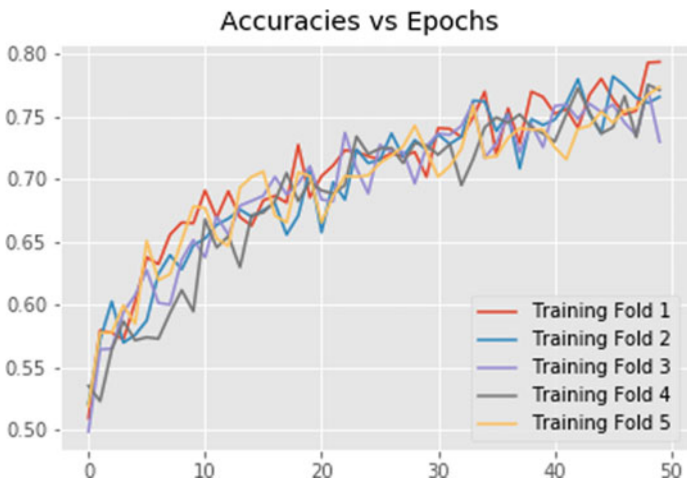


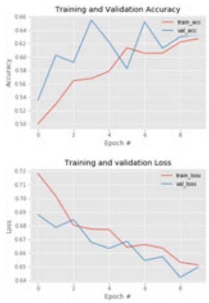
Fig. 4 Accuracy versus epoch for five fold cross validation

Table 4 Comparison of results

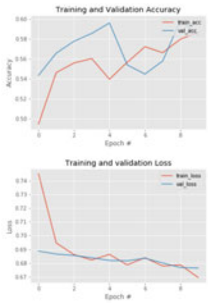
Models	Train acc	Train loss	Test acc	Test loss
CNN	0.80	0.4505	0.78	0.3739
VGG19	0.5830	0.6700	0.5825	0.6790
VGG16	0.6290	0.6510	0.6310	0.6500
InceptionV3	0.7100	0.5500	0.6530	0.6460
DenseNet121	0.7600	1.100	0.5490	0.5400
MobileNetV2	0.8400	0.3800	0.5490	1.10
ResNet50	0.7879	0.4493	0.4724	0.7093

4.2 Comparison of Different Models

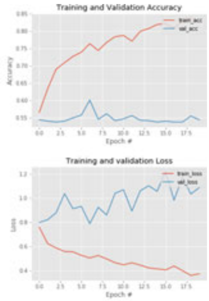
We have trained our dataset using some pretrained model (VGG16, VGG19, ResNet50, MobileNetV2, InceptionV3, DenseNet121) in the purpose of showing



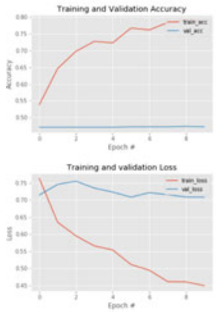
VGG-16 acc and loss



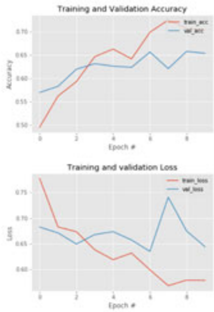
VGG-19 acc and loss



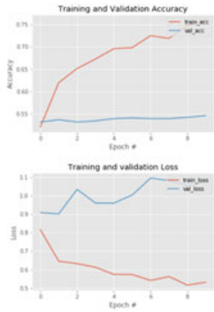
MobileNetV2 acc and loss



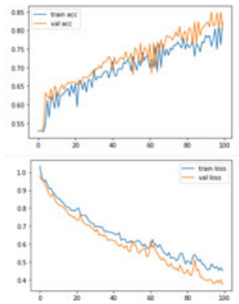
ResNet50 acc and loss



InceptionV3 acc and loss



DenseNet121 acc and loss



Proposed model acc and loss

Fig. 5 Accuracy and loss curve of proposed model and other traditional models

the comparison and also for achieving the best accuracy. After implementing all this models, we have observed that out proposed CNN model outperforms the other models. Table 4 illustrates the training and testing accuracy as well as the training and testing loss of the other models.

In Fig. 5, the curve of accuracy and loss function of our models as well as other models are illustrated.

5 Conclusion

The main objective of this research is to detect fractured humerus from X-ray image using deep learning algorithm. This study exhibits the probability of future use of deep learning algorithm in the filed of orthopedic surgery. This automated system of detecting humerus fractures has some advantages, for example, quick detection, efficiency, and time saving process. However, we have utilized convolutional neural network (CNN) model as a deep learning approach for training our dataset. The X-ray image of fractured and non-fractured humerus bone was used to perform the experiment where total of 1266 X-ray images have been used. Moreover, data augmentation has been applied to increase the size of the dataset which helps to overcome the over-fitting problem. Finally after classification, the accuracy that we gained from this research is 78%.

In future, this research aims at collecting more images to train our model. Moreover, we intent to improve our system by applying the BRBES-based adaptive differential evolution (BRBaDE). This can ameliorate the accuracy by conducting parameter and structure optimization [10–13, 18, 24, 25].

References

1. Adams, M., Chen, W., Holcdorf, D., McCusker, M.W., Howe, P.D., Gaillard, F.: Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. J. Med. Imaging Radiat. Oncol. **63**(1), 27–32 (2019)
2. Ahmed, T.U., Hossain, S., Hossain, M.S., ul Islam, R., Andersson, K.: Facial expression recognition using convolutional neural network with data augmentation. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 336–341. IEEE (2019)
3. Basnin, N., Nahar, L., Hossain, M.S.: An integrated CNN-LSTM model for micro hand gesture recognition. In: International Conference on Intelligent Computing & Optimization, pp. 379–392. Springer, Berlin (2020)
4. Basnin, N., Sumi, T.A., Hossain, M.S., Andersson, K.: Early detection of Parkinson's disease from micrographic static hand drawings. In: International Conference on Brain Informatics, pp. 433–447. Springer, Berlin (2021)
5. Chung, S.W., Han, S.S., Lee, J.W., Oh, K.S., Kim, N.R., Yoon, J.P., Kim, J.Y., Moon, S.H., Kwon, J., Lee, H.J., et al.: Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. **89**(4), 468–473 (2018)
6. De Vries, B., Hegeman, J., Nijmeijer, W., Geerdink, J., Seifert, C., Groothuis-Oudshoorn, C.: Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis. Osteoporos. Int. **32**(3), 437–449 (2021)
7. Demir, S., Key, S., Tuncer, T., Dogan, S.: An exemplar pyramid feature extraction based humerus fracture classification method. Med. Hypotheses **140**, 109663 (2020)
8. Devana, S.K., Shah, A.A., Lee, C., Gudapati, V., Jensen, A.R., Cheung, E., Solorzano, C., van der Schaar, M., SooHoo, N.F.: Development of a machine learning algorithm for prediction of complications and unplanned readmission following reverse total shoulder arthroplasty. J. Shoulder Elbow Arthroplasty **5**, 24715492211038172 (2021)

9. Islam, M.Z., Hossain, M.S., ul Islam, R., Andersson, K.: Static hand gesture recognition using convolutional neural network with data augmentation. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 324–329. IEEE (2019)
10. Islam, R.U., Hossain, M.S., Andersson, K.: A deep learning inspired belief rule-based expert system. IEEE Access **8**, 190637–190651 (2020)
11. Jamil, M.N., Hossain, M.S., ul Islam, R., Andersson, K.: A belief rule based expert system for evaluating technological innovation capability of high-tech firms under uncertainty. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 330–335. IEEE (2019)
12. Kabir, S., Islam, R.U., Hossain, M.S., Andersson, K.: An integrated approach of belief rule base and deep learning to predict air pollution. Sensors **20**(7), 1956 (2020)
13. Karim, R., Andersson, K., Hossain, M.S., Uddin, M.J., Meah, M.P.: A belief rule based expert system to assess clinical bronchopneumonia suspicion. In: 2016 Future Technologies Conference (FTC), pp. 655–660. IEEE (2016)
14. Negrillo-Cárdenas, J., Jiménez-Pérez, J.R., Cañada-Oya, H., Feito, F.R., Delgado-Martínez, A.D.: Automatic detection of landmarks for the analysis of a reduction of supracondylar fractures of the humerus. Med. Image Anal. **64**, 101729 (2020)
15. Negrillo-Cárdenas, J., Jiménez-Pérez, J.R., Feito, F.R.: Automatic detection of distal humerus features: first steps. In: VISIGRAPP (1: GRAPP), pp. 354–359 (2019)
16. Pranata, Y.D., Wang, K.C., Wang, J.C., Idram, I., Lai, J.Y., Liu, J.W., Hsieh, I.H.: Deep learning and surf for automated classification and detection of calcaneus fractures in CT images. Comput. Methods Programs Biomed. **171**, 27–37 (2019)
17. Progga, N.I., Hossain, M.S., Andersson, K.: A deep transfer learning approach to diagnose covid-19 using X-ray images. In: 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 177–182. IEEE (2020)
18. Rahaman, S., Hossain, M.S.: A belief rule based clinical decision support system to assess suspicion of heart failure from signs, symptoms and risk factors. In: 2013 International Conference on Informatics, Electronics and Vision (ICIEV), pp. 1–6. IEEE (2013)
19. Rezaoana, N., Hossain, M.S., Andersson, K.: Detection and classification of skin cancer by using a parallel CNN model. In: 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 380–386. IEEE (2020)
20. Sasidhar, A., Thanabal, M., Ramya, P.: Efficient transfer learning model for humerus bone fracture detection. Ann. Rom. Soc. Cell Biol. 3932–3942 (2021)
21. Sezer, A., Sezer, H.B.: Convolutional neural network based diagnosis of bone pathologies of proximal humerus. Neurocomputing **392**, 124–131 (2020)
22. Shrader, M.W., Sanchez-Sotelo, J., Sperling, J.W., Rowland, C.M., Cofield, R.H.: Understanding proximal humerus fractures: image analysis, classification, and treatment. J. Shoulder Elbow Surgery **14**(5), 497–505 (2005)
23. Sumi, T.A., Hossain, M.S., Islam, R.U., Andersson, K.: Human gender detection from facial images using convolution neural network. In: International Conference on Applied Intelligence and Informatics, pp. 188–203. Springer, Berlin (2021)
24. Uddin Ahmed, T., Jamil, M.N., Hossain, M.S., Andersson, K., Hossain, M.S.: An integrated real-time deep learning and belief rule base intelligent system to assess facial expression under uncertainty. In: 9th International Conference on Informatics, Electronics & Vision (ICIEV). IEEE Computer Society (2020)
25. Zisad, S.N., Chowdhury, E., Hossain, M.S., Islam, R.U., Andersson, K.: An integrated deep learning and belief rule-based expert system for visual sentiment analysis under uncertainty. Algorithms **14**(7), 213 (2021)
26. Zisad, S.N., Hossain, M.S., Andersson, K.: Speech emotion recognition in neurological disorders using convolutional neural network. In: International Conference on Brain Informatics, pp. 287–296. Springer, Berlin (2020)

Obesity and Mental Health During the COVID-19 Pandemic: Prediction and an Exploration of Their Relationship

Shahriar Rahman Khan, Noor Nafiz Islam, S. M. Tamzid Islam, Syed Rohit Zaman, Md. Rezwan-A-Rownok, and Muhammad Nazrul Islam

Abstract Obesity has become a worldwide problem that has rapidly increased with the advancement of technology. This has become one of the most important reasons which reduced life expectancy within the "modern" world. Overweight and obesity tendency continue to increase both in developed and in developing countries. From pediatric to geriatric individuals, it is common in every age group. In different studies around the world, it has been found that overweight and obesity are growing epidemic health concerns. Studies show obesity results in impaired health and premature death. Moreover, during the COVID-19 pandemic, sedentary lifestyle is leading a lot of individuals to obesity. The COVID-19 pandemic also has a negative impact on the mental health of individuals of the affected countries. In this study, we have evaluated 14 different predictive classification models to find the best precision rates to detect the obesity levels and the possibility of being overweight based on the data collected during the epidemic of COVID-19. In the research, apart from finding the best obesity detection model, we also explored the association between obesity and mental health disorders (anxiety and depression) among Bangladeshi people by analyzing their obesity level with their mental health condition, during the ongoing COVID-19 pandemic.

Keywords Machine learning · Obesity · Mental health · COVID-19 · Chi-square

1 Introduction

Obesity and overweight are defined by the World Health Organization (WHO) as excessive fat buildup in specific body areas that can be hazardous to health [12]. The primary causes of rising obesity are an increase in the consumption of high-fat foods, a decrease in physical activity due to the nature of sedentary jobs, new modes

S. R. Khan (✉) · N. N. Islam · S. M. Tamzid Islam · S. R. Zaman · Md. Rezwan-A-Rownok · M. N. Islam
Department of Computer Science and Engineering, Military Institute of Science and Technology, Mirpur Cantonment, Dhaka 1216, Bangladesh
e-mail: shahriarkhan@cse.mist.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_38

of transportation such as cars, buses, and trains, the use of mobile phones and laptop computers, and an increase in urbanization [16]. Many variables, including genetic, metabolic, behavioral, and environmental effects, combine to cause overweight and obesity. In light of the pace with which obesity is spreading, behavioral and environmental effects are blamed rather than biological changes [30]. Food consumption, nutrient turnover, thermogenesis, and body fat storage all have a role in body weight and composition. It appears that food and physical activity habits have a key role in causing obesity. In turn, susceptibility genes may alter energy expenditure, fuel metabolism, and muscle fiber function as well as appetite or dietary preferences [21]. There are various types of obesity, including monogenic, leptin, polygenic, and syndromic, which are caused by biological risk factors, such as genetic background. Social, psychological, and dietary variables are also treated as risk factors [10]. However, several hypotheses exist that state psycho-social and mental health issues are responsible for obesity [14]. Various studies had been conducted to find obesity influence factors through machine learning [8]. Most studies analyzed the disease and developed tools to calculate a person's obesity level, but such tools are limited to calculating the body mass index, food intake, daily water intake, and so on, while ignoring relevant factors such as family background, smoking, transportation usage type, and so on [8, 9, 23].

Although numerous factors cause obesity, it is often considered to be a result of either excessive food intake or insufficient physical activity. Humans take in energy in the form of protein, carbohydrates, fat, and alcohol. This energy needs to be expended properly because if energy intake is not equal to energy expenditure over a given period, then body weight can change which may lead to obesity [15]. A range of chronic diseases can affect patients with obesity, including cardiovascular disease (CVD), gastrointestinal disorders, type 2 diabetes (T2D), joint and muscular disorders, respiratory problems, as well as psychological issues [12]. Mental health issues have a role in the genesis and maintenance of the overweight and obese status of children, adolescents, and adults [31]. Again, the COVID-19 pandemic has affected the social, economic, political, and public health sectors across the world [25, 27]. Mental health problems like post-traumatic stress disorder (PTSD), major depressive disorders (MDD), and bipolar disorder (BD) are also a growing concern in the nations impacted by the COVID-19 pandemic. Individuals with these mental illnesses have raised the risk of obesity by two to three times [4, 17].

Therefore, the primary objectives of this research are as follows: firstly, to creating a novel dataset to predict obesity; secondly, to exploring the machine learning algorithms to find out the best possible one for predicting obesity levels using the collected dataset; and finally, to explore the correlation between obesity and mental health disorders (anxiety and depression) during the COVID-19 pandemic. The rest of this research article is organized as follows. Section 2 presents the literature review. Dataset collection, features selection, and preprocessing of the collected dataset are discussed in Sect. 3. Next, Sect. 4 presents the development of 14 different predictive models and evaluation study of these models with results. Section 5 briefly illustrates the correlation test between obesity level, anxiety, and depression. Finally, Sect. 6 contains the conclusion and future expansion.

2 Literature Review

Obesity has become a research topic of interest, with many studies focusing on the factors that cause the disease. The ability to predict childhood obesity will help early prevention. In [2], the author introduced the use of data mining on childhood obesity prediction, described current efforts in the area, and provided the strength and weaknesses of each technique. Cervantes and Palacio in [6] created an intelligent method based on supervised and unsupervised data mining methods, simple K-means, decision trees (DT), and support vector machines (SVMs) to identify obesity levels and enable consumers and health professionals live a healthy lifestyle in the face of this worldwide crisis. Langarizadeh and Moghbeli [20] highlighted evidence about the application of Naive Bayesian networks (NBNs) in predicting disease, and it tries to show NBNs as the fundamental algorithm for the best performance in comparison with other algorithms. Abdullah et al. [1] showed a children obesity classification of the VI graded school students from different districts of Malaysia. The classification techniques, namely Bayesian network, decision tree, neural networks, and support vector machine (SVM), were implemented and compared on the datasets having 4245 data. This paper evaluates numerous feature selection approaches based on various classifiers [1].

Mental health is one of the factors that lead to obesity. After the COVID-19 pandemic, the social, economic, political, and public health sectors across the world have been affected. Mental health issues such as post-traumatic stress disorder (PTSD), major depressive disorders (MDD), and bipolar disorder (BD) are becoming more prevalent in countries affected by the COVID-19 pandemic [4, 17]. In [3], a survey was done on a total of 589 patients and 40% of them stated that they had begun or escalated drug use to deal with the COVID-19 pandemic as they were facing at least one mental health condition including anxiety or sadness (30.9%), symptoms of a trauma or stress related disorder (26.3%), and signs of a trauma or stress related disorder (26.3%). During the COVID-19 pandemic, mental illnesses like loneliness, depression, anxiety, and sleep disturbance were shown to be common in 71%, 38%, 64%, and 73% of individuals, respectively [7]. The COVID-19 pandemic and the ensuing economic downturn have a severe impact on many people's mental health and created additional hurdles for those who already suffer from mental illness or substance abuse problems. Approximately 4 in 10 individuals in the United States have reported anxiety or depressive symptoms during the pandemic [26].

Again, some studies were conducted to find a relation between mental health and obesity during the COVID-19 pandemic around the world. Moonajilin et al. [24] conducted a cross-sectional survey to find the relationship between obesity and mental health disorders on Bangladeshi adolescents before the COVID-19 pandemic. The findings of the study showed that the prevalence of depression and anxiety was more in adolescents with obesity compared to non-obese adolescents. However, they did not find any significant association between obesity and mental health disorders in adolescents of Bangladesh. Almandoz et al. [3] showed that the COVID-19 pandemic has continued to have a severe influence on the mental health of persons who are obese.

In summary, this literature study raises a few critical issues. First, it is clear that there is no suitable dataset available to predict obesity that was gathered during the COVID-19 epidemic. Second, some of the studies analyzed the disease and developed tools to calculate the obesity level of a person, but such tools are limited to the calculation of the body mass index, food intake, daily water intake, etc., while omitting relevant factors such as family background, smoking, and transportation usage type. Third, no works have been done, where 14 machine learning models were explored and evaluated using a proper dataset to find out the best predictive model that can predict obesity, based on several evaluation metrics, i.e., accuracy, precision, recall, f_1 score, and kappa. Finally, there is not enough work done to find a correlation between obesity and mental health issues (anxiety and depression) during the COVID-19 pandemic. In an attempt to bring together these key points into a solution, the focus of our research has been to improve on those areas to mitigate these problems to some extent.

3 Data Acquisition

3.1 Questionnaires Preparation and Data Collection

The dataset was created through a survey. A total of 30 questionnaires were prepared that include as follows: (a) 14 questions related to obesity detection, (b) 9 questions were adopted from PHQ-9 (The Nine-Item Patient Health Questionnaire) as suggested by Kroenke et al. [18] focusing the screening test of depression, and (c) the remaining 7 questions were adopted from GAD-7 (The Seven-Item Generalized Anxiety Disorder) as suggested by Rutter and Brown [28] focusing on the screening test of anxiety.

A number of characteristics were included in the questionnaires used to construct the dataset. Genders, age, height, weight, location, family obesity history, fast food intake, fruit and vegetable consumption, number of main meals, junk food consumption, soft drink consumption, smoking, daily water consumption, physical activity, use of technology, modes of transportation, mental-emotional disorders, and diagnosed hypertension were all significant in predicting obesity status in adults using ML methods. These risk variables were divided into several categories. The attributes or features related to the dataset include rate of eating of high caloric food, rate of intake of vegetables and fruits, number of major meals, intake of food between meals, daily water intake, soft drink consumption, and smoking. The physical condition traits or aspects are calorie consumption tracking, physical activity frequency, and time spent utilizing technological gadgets. Other factors retrieved were as follows: gender, age, location, height, and weight, as well as the mode of transportation utilized while traveling.

Using height and weight, the initial obesity level or body mass index (BMI) was determined. According to World Health Organization (WHO), the body mass index

(BMI) below 18.5 indicates as underweight, 18.5–24.9 indicates as normal, 25.0–29.9 indicates as overweight, 30.0–34.9 indicates as obesity I, 35.0–39.9 indicates as obesity II, and higher than 40 indicates as obesity III [32].

Based on the answers to the questions for anxiety [29], some scores are given. The GAD-7 score is computed by giving scores of 0, 1, 2, and 3 to the response categories of "not at all", "several days", "more than half the days", and "almost every day", respectively, and summing the values for the seven items. Scores of 5, 10, and 15 are used as the cut-off marks for mild, moderate, and severe anxiety, respectively, and are considered a characteristic of the dataset called the anxiety level.

Similarly, based on the answers to the questions for depression [18], some scores are given. The PHQ-9 score is determined by assigning scores of 0, 1, 2, and 3 to the response categories of "not at all", "several days", "more than half the days", and "almost every day", respectively, and summing the values for the nine items. Scores of 5, 10, 15, and 20 denoted mild, moderate, moderately severe, and severe depression, respectively, and were regarded as a characteristic of the dataset known as the depression level.

Identifying these risk variables might help health officials create or alter existing policies to better control chronic illnesses, particularly those related to obesity risk factors. Obesity level was chosen as the target variable since our primary goal is to develop the best classification model that can properly predict obesity levels, and obesity level, anxiety level, and depression level were chosen to find out the correlation between obesity, anxiety, and depression during the COVID-19 pandemic.

3.2 Respondent's Profile

Data was collected by using a Web platform where anonymous users answered different questions. A total of 304 people (176 male and 128 female) of different age groups ranging from 18 to 55 participated in the survey. Among them, 206 were undergraduate students, 74 serving at different government and non-government organizations; and the remaining 26 were unemployed since starting the COVID-19 pandemic. The dataset was prepared with a view to utilizing it for estimation of the possibility of obesity of individuals.

3.3 Data Preprocessing

The collected survey data was preprocessed to handle missing values, noise removal, and outlier removal. The obesity level column contained 5 types of categories or classes: underweight, normal, overweight, obesity I, obesity II, and higher than 40. So, the class distribution of the obesity level was not equal or close to equal and is instead biased or skewed. Since the dataset also had the class imbalance problem [5], this problem was overcome using synthetic minority oversampling technique

(SMOTE) [13]. This is a statistical methodology used to balance the number of instances in the dataset. The module generates new instances depending on current minority cases supplied as input, while the number of majority cases remains constant as a result of SMOTE's deployment [13].

4 Predictive Model Development

14 machine learning algorithms were used to create ML models to predict obesity levels, that includes: Light Gradient Boosting Machine, Gradient Boosting Classifier, Extreme Gradient Boosting, CatBoost Classifier, Ada Boost Classifier, Decision Tree Classifier, K Neighbors Classifier, Random Forest Classifier, Ridge Classifier, Linear Discriminant Analysis, Logistic Regression, Extra Trees Classifier, SVM-Linear Kernel, Quadratic Discriminant Analysis, and Naive Bayes. The tenfold cross-validation technique was applied for each predictive model. The performance of these models was evaluated based on accuracy, precision, recall, F_1 score, and kappa metrics as shown in Table 1.

Table 1 Accuracy, recall, precision, F_1 score, and kappa of different classifiers for obesity prediction (tenfold cross-validation, sorted by accuracy)

Model	Accuracy	Recall	Precision	F_1 score	Kappa
Light gradient boosting machine	0.8459	0.6524	0.8352	0.8335	0.7478
Gradient boosting classifier	0.8351	0.653	0.8247	0.8228	0.7285
Extreme gradient boosting	0.8135	0.6523	0.806	0.7988	0.6902
CatBoost classifier	0.8027	0.5793	0.7872	0.783	0.6701
AdaBoost classifier	0.7432	0.5475	0.7375	0.7313	0.5794
Decision tree classifier	0.7243	0.5375	0.8022	0.7428	0.5881
K neighbors classifier	0.7054	0.5251	0.6913	0.6822	0.5087
Random forest classifier	0.6838	0.4306	0.6474	0.6503	0.4604
Ridge classifier	0.6514	0.344	0.5575	0.5875	0.3674
Linear discriminant analysis	0.6514	0.3602	0.5792	0.5984	0.3854
Logistic regression	0.6486	0.3716	0.5786	0.5995	0.3751
Extra trees classifier	0.6054	0.3253	0.5459	0.5602	0.2998
SVM-linear kernel	0.4595	0.237	0.3573	0.3701	0.1278
Quadratic discriminant analysis	0.3541	0.29	0.4289	0.2953	0.1087
Naive Bayes	0.1919	0.2451	0.4033	0.1815	0.0936

Table 2 Best predictive model concerning the evaluation metrics

Metrics	Maximum value	Best predictive model
Accuracy	0.8459	Light gradient boosting machine
Recall	0.653	Gradient boosting classifier
Precision	0.8352	Light gradient boosting machine
F_1 score	0.8335	Light gradient boosting machine
Kappa	0.7478	Light gradient boosting machine

4.1 Evaluation Study

The best predictive model concerning the evaluation metrics and values is shown in Table 2. From the evaluation, we found that, with a score of 0.8459, the light gradient boosting classifier model has the highest accuracy of all of the 14 predictive machine learning models mentioned in Table 1, which means this classifier model is approx. 84% accurate. The light gradient boosting classifier model has the highest recall of all of the 14 predictive machine learning models listed in Table 1, with a value of 0.8352. Our model's recall is a statistic that measures how effectively it can find relevant data. We call it sensitivity or the true positive rate. There is a possibility that a person is obese, but our model indicated that the individual is not obese.

Among all the 14 predictive machine learning models shown in Table 1, gradient boosting classifier model has the maximum precision of 0.653, which means when it predicts the obesity level of a person, it is correct around 65% of the time. Precision also provides us with a count of the important data points. We must not start treating someone who does not have obesity but were projected to have it by our model.

The light gradient boosting classifier model has the highest recall of all of the 14 predictive machine learning models listed in Table 1, with a value of 0.8335. With a score of 0.7478, the light gradient boosting classifier model has the highest kappa of all of the 14 predictive machine learning models mentioned in Table 1. Landis and Koch [19] proposed a method for describing the values of kappa metrics. According to their system, a value of 0 indicates that the classifier is ineffective, 0.21–0.40 indicates fair, 0.41–0.60 indicates moderate, 0.61–0.80 indicates considerable, and 0.81–1 indicates practically flawless.

5 Exploring Correlations

Another purpose of this research was to find out the correlation between obesity, anxiety, and depression during the COVID-19 pandemic.

So finally after the dataset was properly preprocessed, 3 categorical columns were generated. They are anxiety, depression, and BMI classification. Anxiety contains 4 attributes: minimal, mild, moderate, and severe. Depression contains 3 attributes:

normal, depression, and severe depression. Similarly, the BMI classification contains 5 attributes: underweight, normal, overweight, obesity I, obesity II, and higher than 40. Correlation tests among the three features (weight, anxiety, and depression) are checked using the chi-square test [11, 22].

5.1 *Obesity Versus Anxiety*

For calculating the correlation between obesity level and anxiety level, the pivot table is prepared from the dataset which is shown in Fig. 1, where all data are considered as the observed values. The expected values are then computed and saved in another pivot table using the observed data from Fig. 1.

$$\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{1}$$

The outcome is 100.95 after applying the chi-square test formula (1), where O is the observed value and E is the expected value. The chi-square table yields a result of 25.00 since the degree of freedom is

$$(m-1) * (n-1) = (5 * 3) = 15,$$

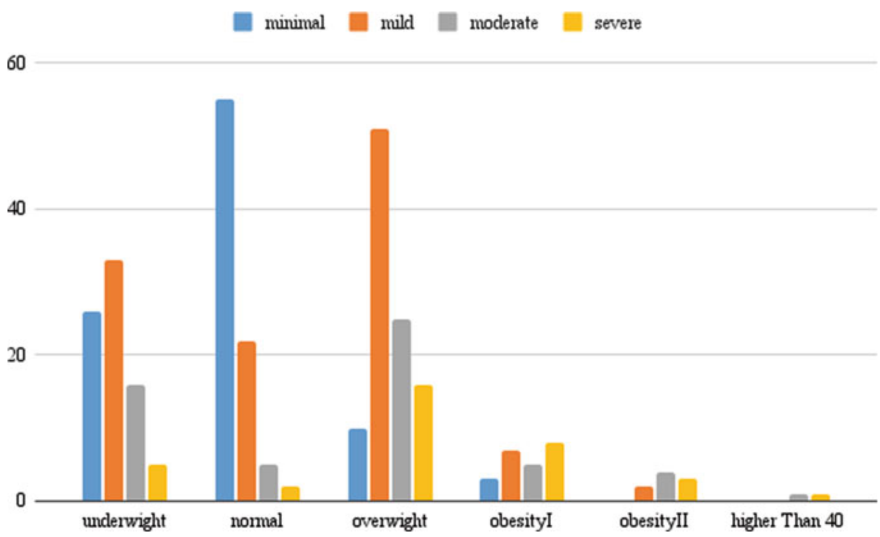


Fig. 1 Observed values of BMI level and anxiety level

and the level of significance is 0.05. According to this step, obesity and anxiety levels are correlated during the COVID-19 pandemic since 100.95 > 25.00 where 100.95 is the chi-square value (X^2), gained from the formula (1) and 25.00 is the chi-square value obtained from the chi-square table.

5.2 Obesity Versus Depression

For calculating the correlation between obesity level and depression level, the pivot table is prepared from the dataset which is shown in Fig. 2, where all data are considered as the observed values. The expected values are then computed and saved in another pivot table using the observed data from Fig. 2.

The outcome is 86.22 after applying the chi-square test formula (1), where O is the observed value and E is the expected value. The chi-square table yields a result of 18.31 since the degree of freedom is

$$(m - 1) * (n - 1) = (5 * 2) = 10,$$

and the level of significance is 0.05.

Similarly, obesity and depression levels are correlated during the COVID-19 pandemic since 86.22 > 18.31 where 86.22 is the chi-square value (X^2), gained from the formula (1), and 18.21 is the chi-square value obtained from the chi-square table.

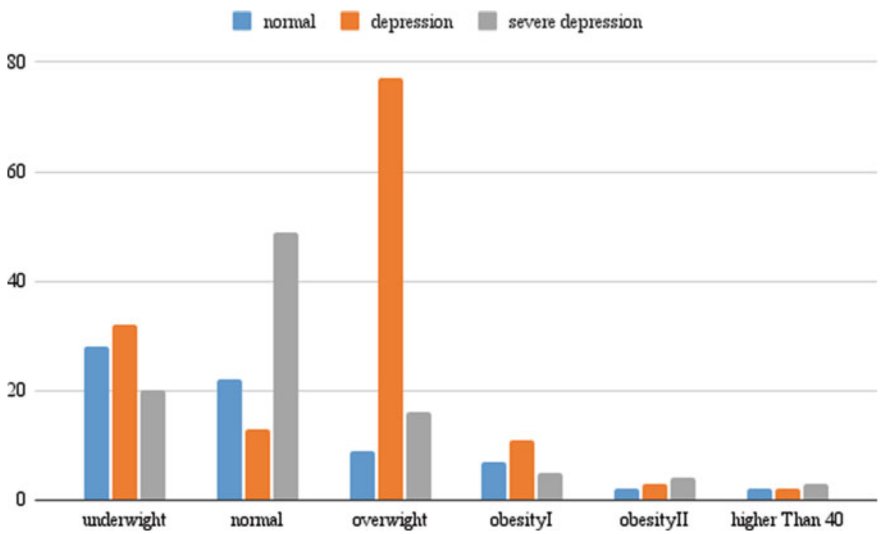


Fig. 2 Observed values of BMI level and depression level

6 Conclusion

Obesity is a disease that affects individuals all over the world, regardless of their age, socioeconomic status, or cultural background. Since 1975, global obesity has nearly quadrupled. In 2016, approximately 1.9 billion individuals aged 18 and above were overweight. Over 650 million of these people were obese. Several methods and solutions to identify or anticipate the sickness have been created to aid in the battle against this disease.

In our work, we collected and preprocessed data during the COVID-19 pandemic and assessed 14 different predictive classification models utilizing that data to reach the highest precision rates in detecting obesity levels and the probability of becoming overweight. In addition to identifying the optimal obesity detection model, we investigated the relationship between overweight/obesity and mental health issues (anxiety and depression) among Bangladeshi individuals during the ongoing COVID-19 epidemic. Among the 14 classification models, we discovered that light gradient boosting machine (LightGBM) is the highest predictive machine learning classifier model. During the COVID-19 outbreak, we also observed links between morbidly obese and mental health disorders (anxiety and depression).

There are a few limitations to this study. To begin, the dataset was acquired in Bangladesh. In future, the information can be gathered from many nations to produce a generic dataset that can aid prediction models in predicting obesity and discovering links between obesity and mental health issues regardless of country, region, weather, gender, dietary habits, and so on. Another restriction is that we have only worked with two types of mental health disorders: anxiety and depression. In future, we will look at additional mental health illnesses that may be linked to obesity during the COVID-19 pandemic.

References

1. Abdullah, F.S., Abd Manan, N.S., Ahmad, A., Wafa, S.W., Shahril, M.R., Zulaily, N., Amin, R.M., Ahmed, A.: Data mining techniques for classification of childhood obesity among year 6 school children. In: International Conference on Soft Computing and Data Mining, pp. 465–474. Springer, Berlin (2016)
2. Adnan, M.H., Husain, W., Rashid, N.A., Hassan, M.F.: Preliminary analysis to investigate accuracy of data mining for childhood obesity and overweight predictions. Adv. Sci. Lett. **24**(10), 7529–7533 (2018)
3. Almandoz, J.P., Xie, L., Schellinger, J.N., Mathew, M.S., Bismar, N., Ofori, A., Kukreja, S., Schneider, B., Vidot, D., Messiah, S.E.: Substance use, mental health and weight-related behaviours during the covid-19 pandemic in people with obesity. Clin. Obes. **11**(2), e12440 (2021)
4. Avila, C., Holloway, A.C., Hahn, M.K., Morrison, K.M., Restivo, M., Anglin, R., Taylor, V.H.: An overview of links between obesity and mental health. Curr. Obes. Rep. **4**(3), 303–310 (2015)
5. Brownlee, J.: A gentle introduction to imbalanced classification (Jan 2020). https://machinelearningmastery.com/what-is-imbalanced-classification/

6. Cervantes, R.C., Palacio, U.M.: Estimation of obesity levels based on computational intelligence. Inf. Med. Unlocked **21**, 100472 (2020)
7. Das, R., Hasan, M.R., Daria, S., Islam, M.R.: Impact of covid-19 pandemic on mental health among general Bangladeshi population: a cross-sectional study. BMJ Open **11**(4), e045727 (2021)
8. DeGregory, K., Kuiper, P., DeSilvio, T., Pleuss, J., Miller, R., Roginski, J., Fisher, C., Harness, D., Viswanath, S., Heymsfield, S., et al.: A review of machine learning in obesity. Obes. Rev. **19**(5), 668–685 (2018)
9. Dugan, T.M., Mukhopadhyay, S., Carroll, A., Downs, S.: Machine learning techniques for prediction of early childhood obesity. Appl. Clin. Inf. **6**(3), 506 (2015)
10. Flegal, K.M., Graubard, B.I., Williamson, D.F., Gail, M.H.: Cause-specific excess deaths associated with underweight, overweight, and obesity. JAMA **298**(17), 2028–2037 (2007)
11. Franke, T.M., Ho, T., Christie, C.A.: The chi-square test: often used and more often misinterpreted. Am. J. Eval. **33**(3), 448–458 (2012)
12. Fruh, S.M.: Obesity: risk factors, complications, and strategies for sustainable long-term weight management. J. Am. Assoc. Nurse Pract. **29**(S1), S3–S14 (2017)
13. Gosain, A., Sardana, S.: Farthest smote: a modified smote approach. In: Computational Intelligence in Data Mining, pp. 309–320. Springer, Berlin (2019)
14. Hach, I., Ruhl, U.E., Klose, M., Klotsche, J., Kirch, W., Jacobi, F.: Obesity and the risk for mental disorders in a representative German adult sample. Eur. J. Public Health **17**(3), 297–305 (2007)
15. Hill, J.O., Wyatt, H.R., Peters, J.C.: Energy balance and obesity. Circulation **126**(1), 126–132 (2012)
16. Hruby, A., Hu, F.B.: The epidemiology of obesity: a big picture. Pharmacoeconomics **33**(7), 673–689 (2015)
17. Islam, M.N., Khan, S.R., Islam, N.N., Rezwan-A-Rownok, M., Zaman, S.R., Zaman, S.R.: A mobile application for mental health care during covid-19 pandemic: development and usability evaluation with system usability scale. In: International Conference on Computational Intelligence in Information System, pp. 33–42. Springer, Berlin (2021)
18. Kroenke, K., Spitzer, R.L., Williams, J.B.: The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. **16**(9), 606–613 (2001)
19. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 159–174 (1977)
20. Langarizadeh, M., Moghbeli, F.: Applying Naive Bayesian networks to disease prediction: a systematic review. Acta Inf. Med. **24**(5), 364 (2016)
21. Martinez, J.A.: Body-weight regulation: causes of obesity. Proc. Nutr. Soc. **59**(3), 337–345 (2000)
22. McHugh, M.L.: The chi-square test of independence. Biochem. Med. Biochem. Med. **23**(2), 143–149 (2013)
23. Montañez, C.A.C., Fergus, P., Hussain, A., Al-Jumeily, D., Abdulaimma, B., Hind, J., Radi, N.: Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2743–2750. IEEE (2017)
24. Moonajilin, M.S., Rahman, M.E., Islam, M.S.: Relationship between overweight/obesity and mental health disorders among Bangladeshi adolescents: a cross-sectional survey. Obes. Med. **18**, 100216 (2020)
25. Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., Agha, R.: The socio-economic implications of the coronavirus pandemic (covid-19): a review. Int. J. Surg. **78**, 185–193 (2020)
26. Nirmita Panchal, R.K.: The implications of covid-19 for mental health and substance use (July 2021). https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/
27. Ozili, P.K., Arun, T.: Spillover of covid-19: impact on the global economy. Available at SSRN 3562570 (2020)

28. Rutter, L.A., Brown, T.A.: Psychometric properties of the generalized anxiety disorder scale-7 (GAD-7) in outpatients with anxiety and mood disorders. J. Psychopathol. Behav. Assess. **39**(1), 140–146 (2017)
29. Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B.: A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med. **166**(10), 1092–1097 (2006)
30. Stein, C.J., Colditz, G.A.: The epidemic of obesity. J. Clin. Endocrinol. Metab. **89**(6), 2522–2525 (2004)
31. Talen, M.R., Mann, M.M.: Obesity and mental health. Primary Care Clin. Off. Pract **36**(2), 287–305 (2009)
32. Weir, C.B., Jan, A.: BMI classification percentile and cut off points (2019)

Deep Learning-Based Skin Disease Detection Using Convolutional Neural Networks (CNN)

Md. Sazzadul Islam Prottasha, Sanjan Mahjabin Farin, Md. Bulbul Ahmed, Md. Zihadur Rahman, A. B. M. Kabir Hossain, and M. Shamim Kaiser

Abstract Skin disease is a common health condition of the human body that greatly affects people's life. Early and accurate disease diagnosis can help the patients in applying timely treatment thus resulting in quick recovery. Recent developments in deep learning-based convolutional neural networks (CNN) have significantly improved the disease classification accuracy. Motivated from that, this study aimed to diagnose two types of skin diseases—Eczema and Psoriasis using deep CNN architectures. Five different state-of-the art CNN architectures have been used and their performance has been analyzed using 10 fold cross validation. A maximum validation accuracy of 97.1% has been achieved by Inception ResNet v2 architecture with Adam optimizer. The performance matrices result imply that, the model performs significantly well to diagnose skin diseases. Additionally, the study demonstrates two approaches for the practical application of the implemented model. (i) Smartphone oriented approach: It integrates the CNN models with the mobile application, (ii) Web server oriented approach: It integrates the CNN model with a web server for real-time skin disease classification.

Keywords Skin disease diagnosis · Dermatology · Psoriasis · Eczema · Deep learning · CNN

1 Introduction

Skin is one of the most prominent and largest organs in the body which supports survival and acts as a barrier against injuries, heat and any form of damage that can be caused by Ultra Violate ray. Unfortunately, over 900 million people are affected

Md. Sazzadul Islam Prottasha (✉) · S. Mahjabin Farin
Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh
e-mail: sazzadulislamprottasha@gmail.com

Md. Bulbul Ahmed · Md. Zihadur Rahman · A. B. M. Kabir Hossain
Bangladesh Army University of Engineering and Technology, Natore, Bangladesh

M. Shamim Kaiser
Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_39

by various types of skin diseases. Therefore, one of the most prevailing diseases in the recent world is skin disease. There exists a variety of skin disorders that can be trouble-free or hurting. Some have situational causes, i.e., environment pollution, while others may be genetic, some skin conditions are trivial and others can be life-threatening. Given the variety of disease incidence variables, proper identification of skin diseases can be critical. Although the skin diseases are not always curable, the treatments aim to reduce the symptoms. The unevenness, tone, presence of hair, and other characteristics of skin make it difficult to examine and analyze the skin disease properly. Skin cancer is one of the most common cancers today, yet it is also avoidable in the majority of cases. Early and accurate diagnosis of skin diseases is crucial for preventing the disease severity. Recent developments in deep learning-based CNN models have significantly improved the classification process. Hence in this study, we aim to diagnose two common types of skin diseases using deep learning methods.

Atopic dermatitis (eczema), a very common skin condition in children, is a condition that makes the skin red and itchy [1]. It is a chronic disease which may be accompanied by asthma or hay fever and tends to flare periodically. Psoriasis, another common skin disorder, which affects the outside of the elbows, knees or scalp, though it can appear on any location [2]. Some people have expressed psoriasis to be itchy, stinging along with burning sensation. Sometimes it is accompanied by other health conditions, such as diabetes, heart disease, and depression. In this paper, we have taken 2 common diseases into consideration, Psoriasis and Eczema to identify their features and then perform classification.

2 Literature Review

The incidence of skin disease is growing in humans tremendously. Over the years different researchers have proposed several machine leaning and deep learning methods and models for detecting skin diseases. Abbadi et al. [3] proposed a system for diagnosing Psoriasis based on the skin color and texture features derived from Gray Level Co-occurrence Matrix (GLCM). Feed-forward neural networks has been used to classify input images among the infected and non-infected Psoriasis. The work by Połap et al. [4] proposed a sensor-based intelligent skin disease detection system that can identify the skin diseases by simply analyzing the footage from camera. Their proposed CNN architecture shows a promising result with admissible training accuracy of 82.4%. Considering the skin disease diagnosis process is costly and time-consuming, an automatic eczema detection and severity measurement model is proposed in [5]. Based on the image color and texture feature, the proposed system can detect the infected areas of eczema. Support Vector Machine (SVM) classifier is used to classify the identified regions as mild or severe. Similar to this, various machine learning classifiers have been used for identifying chronic kidney disease at [6]. The experiment was performed on reduce chronic kidney disease dataset where

Naive Bayes achieved the highest classification accuracy of 99.1%. The experiment was conducted on reduced dataset with different numbers of attributes. Considering multiple criterion, the dataset attributes were removed which eventually increased the classification accuracy.

A cloud computing-based architecture has been proposed by Abdulbaki et al. [7]. In this paper, along with cloud computing, Back-propagation Neural Network (BpNN) has been combined to diagnose the eczema. Janoria et al. [8] proposed a transfer learning-based CNN architecture for identifying various skin cancer. They used CNN for extracting the features from the images while machine learning classifiers has been used for diagnosing the diseases. The experimental analysis shows that VGG-16 CNN model with the K-Nearest Neighbor algorithm achieved the highest accuracy of 99%. The work by Bhadula et al. [9] analyzed the performance of different machine learning algorithms with Convolutional Neural Network (CNN). The CNN model outperformed all of the other machine learning algorithms with a training and testing accuracy of 99.1% and 96% respectively. The work by Shanthi et al. [10] proposed a CNN architecture for automatic diagnosis of skin diseases. The proposed model consists of 11 layers including multiple convolution, pooling, activation layer, and softmax classifier. The experimental analysis shows that the proposed model achieved an accuracy of around 99%. Similar types of CNN model has been proposed for kidney disease [11], heart disease [12], rice plant disease [13], Alzheimer's disease [14] diagnosis.

3 Methodology

In this section, the details of the skin disease diagnosis process along with various CNN models are discussed in appropriate subheadings. Figure 1 illustrates the details of the working procedure. The procedure starts with data collection followed by categorization of the collected images. Then the images are resized and noisy data are discarded. Image augmentation are performed for increasing the dataset. After that, the features are extracted using CNN models and finally the classification is performed.

3.1 Dataset Collection

Skin is one of the visible and largest organs which supports survival but sadly, there are around 900 million people who are affected by different kinds of skin diseases. Some skin conditions are trivial, and others can be life-threatening. In our work we have considered two common types of skin diseases: Atopic dermatitis (Eczema) and Psoriasis. The disease affected skin images has been collected from the resources as follows:

1. DermNet NZ [15]
2. University of IOWA Health Care [16]
3. Dermnet Skin Disease Atlas [15].

After collecting the images we have identified and discarded some noisy, low contrast images from the dataset manually and finally formulated a dataset of 1000 disease affected skin images consisting of 490 Eczema images and 510 Psoriasis images. We collected the sample skin disease images of different body parts, hence we get a fully representative set of images for each of the disease classes. Sample dataset images has been shown in Fig. 2.

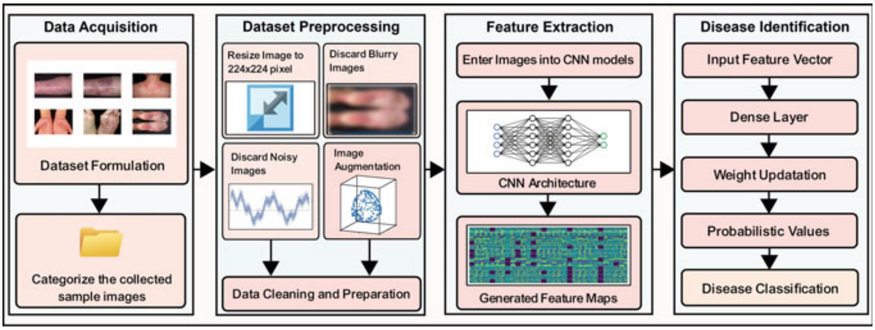


Fig. 1 The working procedure of deep learning-based skin disease prediction model where firstly the data is collected followed by categorizing the dataset. Various data pre-processing and cleaning operations are performed and features are extracted using CNN architectures. Finally, the classification is performed using the dense layer of CNN

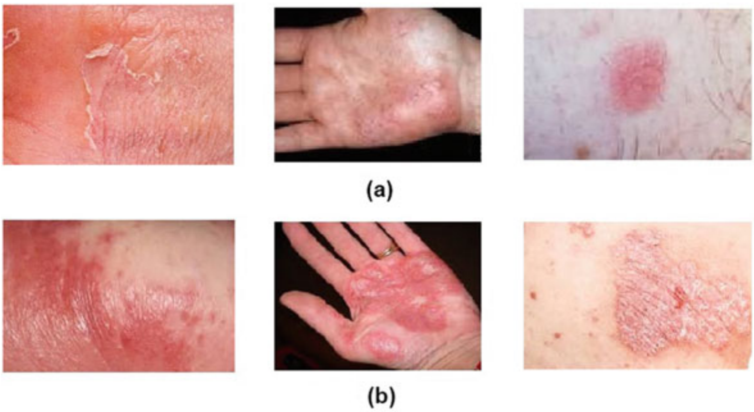


Fig. 2 Sample images of our skin disease dataset, **a** Eczema, **b** Psoriasis

Table 1 Total training and validation images after augmentation process

Disease type	Before augmentation	After augmentation	Training data	Validation data
Eczema	490	2940	2352	588
Psoriasis	510	3060	2448	612
Total	1000	6000	4800	1200

3.2 Data Augmentation

To reduce the over-fitting problem, data augmentation plays an important role that enables practitioners to significantly increase the diversity of data available for training models without collecting new data. Data augmentation techniques such as brightness enhancing by 20%, contrast enhancing by 10%, random rotation of 5 °C and horizontal flipping has been used on our dataset for better interaction and considering all possible variations. In this way, we generated 5 different augmented versions of each sample images. Table 1 shows the number of dataset images after the augmentation process.

We have used 80–20 data split for training and validation data. Thus, we have 4800 training images and 1200 validation images. For testing purpose, we have used 400 skin disease images collected from the Internet. These testing images comprise similar characteristics of the training images. The testing images were totally isolated from the training and validation data for reducing any kind of predilection.

3.3 CNN Architectures

Convolutional Neural Network is supervised machine learning algorithms to specify image identification and classification, trained by using labeled data with their respective classes. In our study, we have used 5 different CNN architectures selected based on several criteria.

VGG-16 [17] architecture which is a very deep CNN model comprising 3×3 filters throughout and consists of large parameter size. The model provided high accuracy in imagenet dataset. The inception v3 [18] architecture is also deep CNN architecture like the VGG-16 but using the inception module the parameter size is reduced significantly yet it provides high accuracy. ResNet-50 [19] architecture has residual connections within the model. The ResNet-50 architecture overcomes the over-fitting problem using that residual connections. MobileNet v2 [20] architecture uses depthwise separable convolutions which reduces the parameter size and accelerates the classification process. This architecture provides substantially high accuracy in imagenet dataset and particularly useful for integrating CNN models with mobile applications because of the small parameter size. Inception ResNet-v2 [21] architecture combines the inception module with the residual connections. This is a improvement over the inception model which provides better accuracy in classifying multiple classes. Therefore, we used our dataset on these architectures to check how it performs under different criterion. We used fine tuning method for training the architectures where we unfreezed the final dense layer of the CNN architectures and freezed all other layer weights. Then we trained the dense layer using our skin disease dataset and initialize random weights to the nodes. The networks adjusted the weights accordingly with the help of optimizers.

3.4 Optimization Algorithms

Optimizers are algorithms used to change the weights and learning rates of CNN architectures to improve the accuracy. In our study, we have used 2 different optimizers namely Adam [22] and Rmsprop [23]. Adam (Adaptive Moment Estimation) works with momentum of first and second order [22]. Adam computes adaptive learning rates for each parameter. Besides storing an exponentially decaying average past gradients like Rmsprop, Adam also stores the average past gradients exponential. The past decaying averages and past squared gradients P_t and Q_t, respectively, are defined in Eqs. 1 and 2.

$$P_t = \beta_1 P_{t-1} + (1 - \beta_1) g_t \tag{1}$$

$$Q_t = \beta_2 Q_{t-1} + (1 - \beta_1) g_t^2 \tag{2}$$

Here, P_t and Q_t represents the estimated mean and uncentered variance of the gradients, respectively. The P_t and Q_t are biased toward zeros because of initial time steps and small decaying rates. Hence the corrected bias is calculated as,

$$\widehat{P_t} = \frac{P_t}{1 - \beta_1^t} \tag{3}$$

$$\widehat{Q_t} = \frac{Q_t}{1 - \beta_2^t} \tag{4}$$

Finally, for updating the weights considering η step size, the weight W_t is calculated as,

$$W_t = W_{t-1} - \frac{\eta}{\sqrt{\widehat{Q_t}} + \varepsilon} \hat{P}_t \tag{5}$$

The default values for β_1, β_2, ε are 0.9, 0.999 and 10^{-8} respectively. This way the Adam optimizer updates the weight of the nodes considering average past gradients.

Rmsprop is another weight updating algorithm which uses a decaying average of partial gradients in the adaptation of the step size for each parameter [23]. The decaying average of partial gradients allows the algorithm to forget early gradients and focus on the most recently observed partial gradients during the progress of the weight updating and helps to converge quickly. The average of the squared gradient for each weight $A(g_t^2)$ is calculated as,

$$A(g_t^2) = 0.9A(g_{t-1}^2) + 0.1\, g_t^2 \tag{6}$$

For updating the weights considering η step size, the weight V_{t+1} is calculated as,

$$V_{t+1} = V_t - \frac{\eta}{\sqrt{A(g_t^2) + \varepsilon}} g_t \tag{7}$$

The default values for η and ε are 0.001 and 10^{-8}, respectively.

4 Results

4.1 Experimental Setup

The experiment was conducted using a machine having a configuration of Intel core i7 processor, 16 GB Ram, Nvidia Gtx-1060 graphics. For performing the classification, Keras framework with tensorflow back-end has been used to train the models in colaboratory. The dataset has been divided into 80–20 split for train, validation, respectively. A total of 400 images has been considered for testing the models. 10 fold cross validation with 100 epochs has been considered during the training and validation process. RmsProp and Adam optimizer algorithm has been used for updating the weights. Since our dataset consists of only 2 classes hence binary cross entropy loss function has been used. The hyper parameters used in our experiment are as follows:

Input Image Size: 224×224
Epoch: 100
learning rate: 0.001
Classification: Softmax
Momentum: 0.9
Optimizer: Adam, RmsProp

4.2 CNN Performance

First of all the input images are fed through the CNN architectures. Each of the CNN model consists of multiple convolution and pooling layer. Each of these layers convert the input images into feature vectors. The feature maps of eczema and psoriasis sample images are shown in Fig. 3.

Fine tuning method has been used for the classification process where we freezed all of the layers of CNN architectures except the final dense layer. In the final dense layer the weight has been updated using Adam and Rmsprop optimizer. The detailed experimental result of various CNN architectures is shown in Table 2.

The result reported in Table 2 indicates that Adam and Rmsprop optimizer both performed similarly in detecting the skin diseases. Adam achieved the highest training, validation, and testing accuracy. However, in overall scenario, Adam and Rmsprop performed identically.

Figure 4 demonstrates the bar graph of training, validation, and testing accuracy of various CNN architectures. From the experimental analysis we observed that, Inception ResNet v2 architecture with Adam optimizer achieved the highest training, validation, and testing accuracy. MobileNet v2 architecture also performed sig-

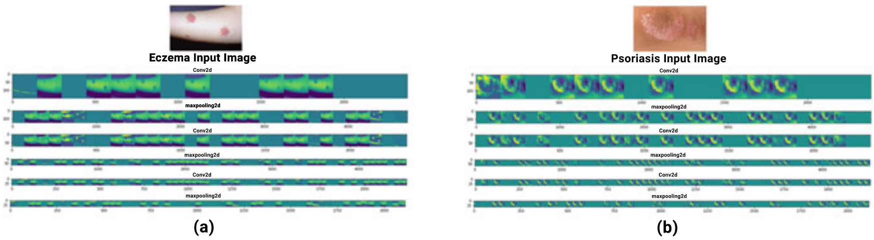


Fig. 3 Feature maps generated from input skin disease image. Here **a** illustrates the original eczema disease image followed by the feature maps generated by multiple convolution and pooling layer, **b** illustrates the original psoriasis disease image followed by the feature maps generated by multiple convolution and pooling layer

Table 2 Accuracy metrics of various CNN architectures using Adam and RmsProp optimizer

Architecture	Training accuracy	Validation accuracy	Testing accuracy	Parameter size (millions)
		Adam optimizer		
VGG-16	98.7	95.8	93.5	134
Inception v3	97.7	91.3	85.6	23
Resnet-50	86.1	72.3	69.8	25
MobileNet v2	99.3	96.3	95.1	**3.5**
Inception ResNet-v2	**99.7**	**97.1**	**95.8**	55
		RmsProp optimizer		
VGG-16	99.4	95.2	91.8	134
Inception v3	98.1	90.5	86.0	23
Resnet-50	87.5	75.8	70.1	25
MobileNet v2	99.2	95.7	94.0	**3.5**
Inception ResNet-v2	99.3	96.5	94.3	55

Here bold font indicates best result

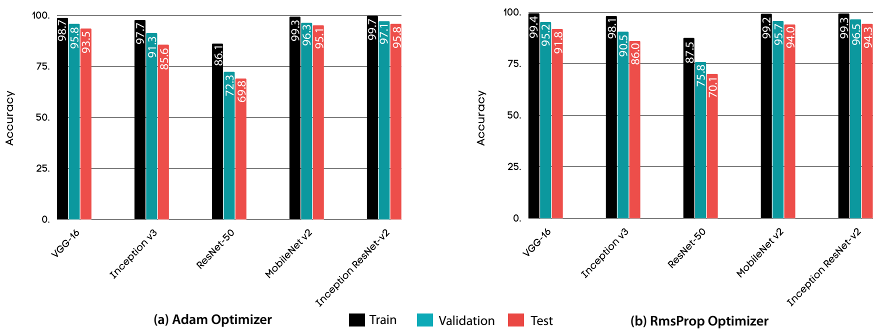


Fig. 4 Accuracy graph of various CNN architectures. Here **a** represents the accuracy graph using Adam optimizer, **b** represents the accuracy graph using RmsProp optimizer

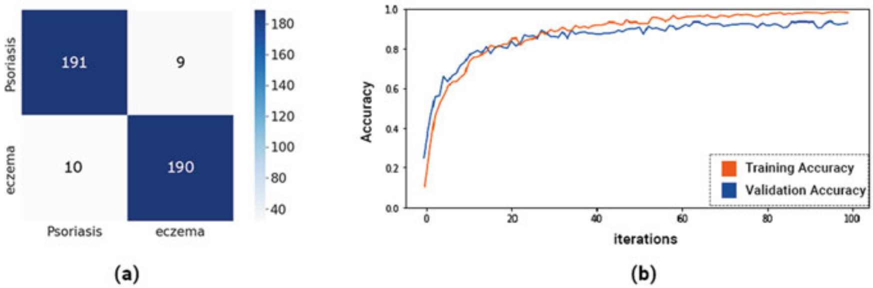


Fig. 5 Inception ResNet v2 architecture performance with Adam optimizer. Here **a** represents the confusion matrix on testing dataset, **b** represents the training and validation accuracy curve

nificantly well while having a smaller parameter size. VGG-16 and Inception v3 both achieved more than 90% training and validation accuracy. ResNet-50 architecture showed over-fitting characteristics with both Adam and Rmsprop optimizer . It achieved high accuracy during the training process, however failed in validation and testing data. Since Inception ResNet v2 architecture with Adam optimizer achieved the highest overall accuracy, the confusion matrix on testing dataset along with accuracy curve of training and validation data using Inception ResNet v2 architecture is demonstrated in Fig. 5.

From Fig. 5a confusion matrix we can see that, Inception ResNet v2 architecture correctly identified the test set with great precision. From Fig. 5b we can see the training and validation curve where the accuracy curve gradually increased and there is no sign of over-fitting characteristics. After 80 iteration the curve reached to an optimum level. Considering the validation and testing accuracy, it is evident that Inception ResNet v2 architecture correctly diagnosed the skin diseases.

Table 3 Comparison of proposed model with existing work

Author	Proposed method	No. of Disease	Performance
Abbadi [3]	GLCM feature extraction with feed forward neural network	1 (psoriasis)	All samples are detected correctly
Srivastava [5]	Neural Network model to identify eczema using texture feature	1 (eczema)	90% accuracy on testing dataset
Bhadula [9]	Machine learning classifiers based approach	3 (acne, lichen planus, sjs ten)	96% accuracy on testing dataset
Shanthi [10]	CNN based classification approach	4 (Acne, Keratosis, Eczema, Urticaria)	Accuracy: 98.6–99.04%
Proposed model	Inception ResNet-v2 architecture combined with Adam optimizer	2 (eczema, psoriasis)	Training: 99.7% Validation: 97.1% Testing: 95.8%

4.3 Discussions

Several researchers have proposed various methodology for detecting different skin diseases. A statistical comparison between various works along with our proposed solution is shown in Table 3. Based on the results reported in Table 3, we can observe that most of the works focused on feature and texture selection process because of the fact that different skin disease comprise different shapes and colors. And most of the CNN methods achieved a similar level accuracy.

From the experimental analysis on skin disease dataset, we have got the following observations:

1. Rather than transfer learning method, fine tuning method can significantly improve the accuracy of any CNN model on a particular dataset.
2. In terms of execution time, Adam optimizer performed faster than the Rmsprop optimizer. Rmsprop optimizer took more time to perform the classification task.
3. Inception ResNet v2 architecture achieved the highest training, validation, and testing accuracy with Adam optimizer.
4. MobileNet v2 architecture achieved similar accuracy of Inception ResNet v2 architecture with 15 times less number of parameters.
5. In terms of accuracy, Adam and Rmsprop optimizer both performed similarly. Hence, both of the optimizer can be used for classification task.
6. ResNet-50 architecture showed over-fitting characteristics with both Adam and Rmsprop optimizer.
7. A learning rate of 0.001 with momentum of 0.9 performs better in updating the weights. However, learning rate of 0.01 and 0.0001 also performed similar.

5 Practical Applications

Skin disease detection is a critical task and requires a series of pathological laboratory tests for the proper diagnosis. Pathological laboratory tests take time to produce findings, which prolongs the disease diagnosis procedure. As a result, we've put together two practical applications for detecting skin diseases quickly and correctly. The suggested framework can be effectively utilized in real-time diagnosing and assessing the severity of the disease.

5.1 *Smartphone Oriented Approach*

Due to the exponential growth in their number and availability worldwide, mobile devices are now one of the most widely used communication and data transfer equipment. Nowadays, practically everyone has a mobile device, so it will be easy for the user to scan the skin images and receive the results instantly. The CNN model can be integrated with the mobile app to easily identify the skin diseases. The mobile application framework is illustrated in Fig. 6.

To implement the application, firstly the CNN models are freezed with the weights updated by the optimizer. After that the freezed CNN model is integrated with the android application. For mobile devices a lightweight model is more preferred, hence MobileNet v2 architecture is more suitable for performing the classification task within the application. When a user takes a picture with the mobile app, the image is compared to the saved CNN model, and the results are presented to the user.

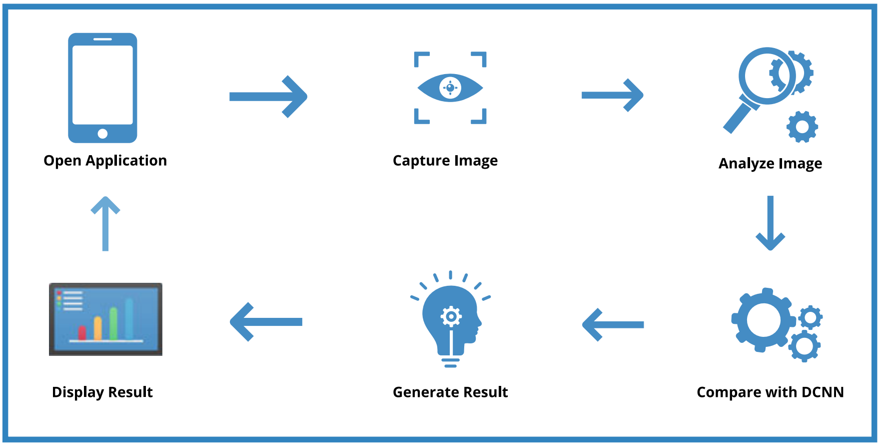


Fig. 6 Smartphone oriented framework for skin disease detection

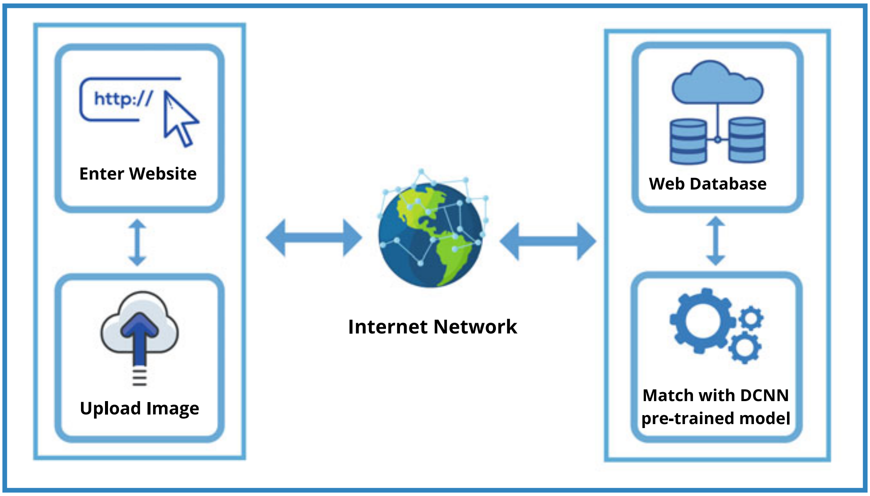


Fig. 7 Web server oriented framework for skin disease detection

5.2 Web Server Oriented Approach

Apart from mobile devices, websites are also used extensively by users because of it's heterogeneity and robustness. Compared to a mobile device, a web server has the ability to store and handle far more complicated data. Hence, web-based skin disease detection system can be a good solution for a quick and accurate disease diagnosis. In mobile devices we can only use lightweight CNN models whereas in web-based server we can use large CNN models, which in terms increase the efficacy of the disease detection process. The proposed web server oriented skin disease detection framework is shown in Fig. 7.

To implement the web server-based detection system, the pre-trained deep CNN models should be saved in the web database. In the front end, users log on to the website and uploads the skin disease images. The CNN model saved in the web database is then compared to the uploaded images. Finally, the matching results are displayed to the users. In this way the users can get the proper diagnosis instantly.

6 Conclusion

Disease recognition along with the help of computational assistance will assist medical science to prosper in a bigger dimension. In this paper, we proposed deep learning-based skin disease detection methods using different CNN architectures to identify 2 common skin diseases named eczema and psoriasis. Fine tuning method was used to train 5 different CNN architectures using Adam and Rmsprop optimizer. Among all

CNN architectures, Inception ResNet v2 architecture outperformed all the other CNN architectures with a validation and testing accuracy of 97.1% and 95.8%, respectively. Finally, two approaches for the practical application has been demonstrated, (i) Smartphone oriented approach and (ii) Web server oriented approach. These practical application can be effectively utilized in real-time diagnosing and assessing the severity of the skin diseases.

References

1. Ruzicka, T., Ring, J., Przybilla, B.: Handbook of atopic eczema, 2nd edn. Springer Science & Business Media (2013)
2. Lowes, M.A., Suarez-Farinas, M., Krueger, J.G.: Immunology of psoriasis. Ann. Rev. Immunol. **32**, 227–255 (2014)
3. Al Abbadi, N.K., Dahir, N.S., AL-Dhalimi, M.A., Restom, H.: Psoriasis detection using skin color and texture features. J. Comput. Sci. **6**(6), 648–652 (2010)
4. Połap, D., Winnicka, A., Serwata, K., Kęsik, K., Woźniak, M.: An intelligent system for monitoring skin diseases. Sensors **18**, 2552 (2018)
5. Alam, M.N., Munia, T.T.K., Tavakolian, K., Vasefi, F., MacKinnon, N., Fazel-Rezai, R.: Automatic detection and severity measurement of eczema using image processing. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1365–1368. IEEE (2016)
6. Khan, M.T.I., Prottasha, M.S.I., Nasim, T.A., Mehedi, A.A., Pranto, M.A.M., Asad, N.A.: Performance analysis of various machine learning classifiers on reduced chronic kidney disease dataset. Int. J. Rec. Res. Rev. (IJRRR), **XIII**(4), 16–21 (2020)
7. Shawkat Abdulbaki, A., Najim, S.A.M., Abdulrasoul Khadim, S.: Eczema disease detection and recognition in cloud computing. Res. J. Applied Sci. **13**, 182–188 (2019)
8. Janoria, H., Minj, J., Patre, P.: Classification of skin disease from skin images using transfer learning technique. In: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 888–895. IEEE (2020)
9. Bhadula, S., Sharma, S., Juyal, P., Kulshrestha, C.: Machine learning algorithms based skin disease detection. IJITEE **9**(2), 4044–4049 (2019)
10. Shanthi, T., Sabeenian, R.S., Anand, R.: Automatic diagnosis of skin diseases using convolution neural network. Microprocess. Microsyst. **76**, 103074 (2020)
11. Bhaskar, N., Manikandan, S.: A deep-learning-based system for automated sensing of chronic kidney disease. IEEE Sens. Lett. **3**, 1–4 (2019)
12. Shankar, V., Kumar, V., Devagade, U., Karanth, V., Rohitaksha, K.: Heart disease prediction using CNN algorithm. SN Comput. Sci. **1**, 170 (2020)
13. Prottasha, M.S.I., Hossain, A.B.M., Rahman, M.Z., Reza, S.M.S., Hossain, D.A.: Identification of various rice plant diseases using optimized convolutional neural network. Int. J. Comput. Dig. Syst. (IJCDS) **9** (2021)
14. Farooq, A., Anwar, S., Awais, M., Rehman, S.: A deep CNN based multi-class classification of Alzheimer's disease using MRI. In: 2017 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6. IEEE, Beijing, China (2017)
15. Dermnet: Dermatology Pictures—Skin Disease Pictures. http://www.dermnet.com/images/Eczema-Photos. Last accessed 30 Aug. 2019
16. Psoriasis Department of Dermatology. https://medicine.uiowa.edu/dermatology/psoriasis. Last accessed 27 Aug. 2019
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR. abs/1409.1556 (2015)

18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. IEEE, Las Vegas, USA (2016)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Las Vegas, USA (2016)
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520. IEEE, Salt Lake City, UT (2018)
21. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence, pp. 4278–4284 (2016)
22. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference for Learning Representations, IEEE, pp. 1–13, San Diego (2015)
23. Hinton, G.: Lecture 6.5—RMSProp. COURSERA: neural networks for machine learning. Technical report, COURSERA (2012)

Automated Detection of Diabetic Foot Ulcer Using Convolutional Neural Network

Pranta Protik, G M Atiqur Rahaman, and Sajib Saha

Abstract Diabetic foot ulcers (DFU) are one of the major health complications for people with diabetes. It may cause limb amputation or lead to life-threatening situations if not detected and treated properly at an early stage. A diabetic patient has a 15–25% chance of developing DFU at a later stage in his or her life if proper foot care is not taken. Because of these high-risk factors, patients with diabetes need to have regular checkups and medications which cause a huge financial burden on both the patients and their families. Hence, the necessity of a cost-effective, re-mote, and fitting DFU diagnosis technique is imminent. This paper presents a convolutional neural network (CNN)-based approach for the automated detection of diabetic foot ulcers from the pictures of a patient's feet. ResNet50 is used as the backbone of the Faster R-CNN which performed better than the original Faster R-CNN that uses VGG16. A total of 2000 images from the Diabetic Foot Ulcer Grand Challenge 2020 (DFUC2020) dataset have been used for the experiment. The proposed method obtained precision, recall, F1-score, and mean average precision of 77.3%, 89.0%, 82.7%, and 71.3%, respectively, in DFU detection which is better than results obtained by the original Faster R-CNN.

Keywords Diabetic foot ulcer · Object detection · Convolutional neural network · Deep learning · Faster R-CNN

1 Introduction

Diabetes mellitus (DM), also known as diabetes, is a serious health condition caused by high blood sugar levels due to the insufficient production of insulin. It can lead to several life-threatening conditions such as kidney failure, cardiovascular diseases,

P. Protik (✉) · G. M. Atiqur Rahaman
Computational Color and Spectral Image Analysis Lab, Computer Science and Engineering Discipline, Khulna University, Khulnaf, 9208, Bangladesh
e-mail: panta1515@cseku.ac.bd

S. Saha
Australian e-Health Research Centre, CSIRO, Perth, WA, Australia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_40

eye diseases [18, 19], and diabetic foot ulcers. Among these, DFU is a serious health complication which in case of late detection and improper treatment may lead to limb amputation. Every year, over one million patients with diabetes lose a portion of their legs as a result of this condition [1]. Figure 1 displays some example images of DFU. The treatment of this disease is a global healthcare problem, and the currently available clinical treatments require high diagnostic costs along with a lengthy treatment process. The treatment procedures involve a thorough evaluation of a patient's medical history as well as careful examination of the wounds by a DFU specialist. Sometimes, it may require some additional tests too like CT scans, X-ray, etc. [9]. Although this type of treatment provides the patients with a powerful and positive outcome, it consumes a lot of time and money and it has a big financial impact on the patient's family. The situation is more awful in the case of developing countries where the treatment charge may get up to 5.7 times the annual income [4]. In this scenario, computer-aided systems may lead to a simple yet effective diagnosis system for DFU. Recently, computer vision and image processing have been successfully applied in several fields of medical applications. So, an expert computer-aided system can expand the availability of automated DFU detection systems among people from all backgrounds while also reducing the huge financial burden they face.

Modern convolutional neural network (CNN)-based object detector, viz. Faster R-CNN [15], has been found to outperform other state-of-the-art systems in locating

Fig. 1 Sample images of diabetic foot ulcers

the region of anomalies in medical images including DFU [3]. Therefore, we have proposed an amended version of the Faster R-CNN in the context of automated DFU. The proposed approach has been found to outperform the original Faster R-CNN on the publicly available DFUC 2020 dataset, which is evident through improved precision, recall, and F1-score.

2 Related Work

The number of diabetic patients is raising day by day, and as a result, the research interest in DFU is also increasing. Several research communities are trying to build an expert system that would assist the treatment procedure of DFU. Still, there are not enough research studies found in this problem domain.

Liu et al. [14] and Saminathan et al. [20] used infrared thermal images to perform automated detection of DFU. They used the temperature characteristic and asymmetry analysis of those images to find out the ulcer area. The temperature distribution in the photos of healthy foot is often symmetrical, but asymmetry is evident in diabetic foot ulcers. Using this characteristic, they extracted features from several parts of the foot and used a support vector machine to categorize the regions into normal and ulcer. Although their work exhibited good results in detecting DFU, they also suffered some drawbacks in case of deformed or amputated feet or if the positioning of feet in images was different.

A mobile app titled MyFootCare was created by Brown et al. [2] in 2017 to help diabetic patients self-monitor their foot condition. The app used the patient's foot images and performed segmentation on them to find out the affected areas. The app was designed to inspire the patients to take a photo of the affected area using their mobile phone so that the image can get analyzed and the patients would get visual feedback on the healing process. However, the system was tested only on three people so its impact on real-world cases is still unknown.

In 2017, an approach to segment diabetic foot ulcer lesions using fully convolutional network (FCN) was proposed by Goyal et al. [12]. A dataset containing 705-foot images was introduced and used for the experiments. They presented a two-tier transfer learning technique for segmenting the ulcer and surrounding skin. From the experiment, they obtained a fairly high accuracy which demonstrates the possibilities of FCNs in the case of diabetic foot ulcer segmentation.

Yap et al. [26] used some deep learning-based state-of-the-art algorithms to detect diabetic foot ulcers from patients' foot images. They used Faster R-CNN, YOLO, EfficientDet, and cascade attention network and compared the outcomes of these methods. Prior to the training of each method, they used data augmentation in order to increase the amount of training images. Their findings demonstrate that the efficiency at which those state-of-the-art methods perform is inadequate for effective DFU detection.

3 Background

The Faster R-CNN [15] which is an evolution of the Fast R-CNN [8] is the most widely used state-of-the-art object detection algorithm. The evolution between the versions was in terms of reduced time, improved efficiency, and performance improvement. The Faster R-CNN network consists of (a) a region proposal network (RPN) which is a CNN that generates bounding boxes for possible objects in the images; (b) a feature generation stage that uses a CNN to generate a feature map; (c) a classification layer to predict the class of the object; and (d) a regression layer to precisely localize the object. Both the region proposal network and the classification layer share the convolutional layers. In Faster R-CNN, the region proposal network is used to help the detection process by proposing some bounding boxes on possible object locations. Later, the classification module uses the proposals to find out where to look for an object and classifies them. Figure 2 describes the whole Faster R-CNN architecture. The Faster R-CNN is a high-powered and one of the best object detection methods available. It achieved the highest accuracy in the PASCAL VOC dataset in 2007 and 2012 [5, 6]. Later, it served as the foundation for the winner models of ImageNet detection and localization in ILSVRC 2015 and COCO detection in the COCO 2015 competition [15]. It may perform slower than some recently developed algorithms, but in terms of detection accuracy, it still outperforms the other methods.

4 Methodology

This section includes a detailed explanation of the DFU dataset as well as the implementation details of our proposed Faster R-CNN for DFU detection.

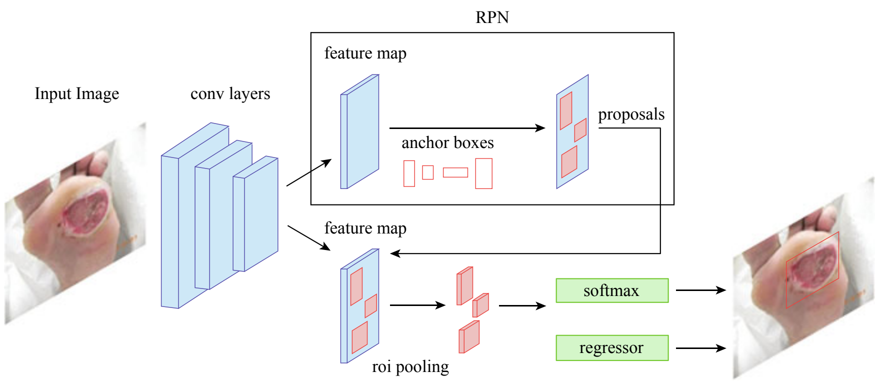


Fig. 2 Faster R-CNN architecture

4.1 DFU Dataset

The images used for the detection of DFU are part of the Diabetic Foot Ulcer Grand Challenge 2020 (DFUC2020) dataset [3, 9–11, 25]. The dataset was made by collecting images of diabetic patients from the Lancashire Teaching Hospital (LTH). Later, the ulcer areas on those images were marked as a rectangular box by the DFU specialists [3] to be used as ground truth. The dataset contains 2000 training images, 200 validation images, and 2000 testing images. To increase the performance and lower the computational costs of the deep learning algorithms, all the images were downsized to 640 × 480 pixels. Figure 3 depicts some sample images of the dataset. The dataset was just recently made publicly available, and as a result, researches based on it are rarely found.

4.2 Data Augmentation

Data augmentation is a popular strategy to increase the diversity of the available training data without actually collecting any new data. It is a common practice to do data augmentation prior to training deep convolutional neural network models [17].

Fig. 3 Examples of the DFUC2020 dataset

In the context of the biomedical field, the process of collecting a large volume of medical images is very costly and time consuming. Hence, in line with other works in the literature, we performed horizontal flip, vertical flip, and rotations by 180° to increase the volume of training images.

4.3 Enhancement of the Original Faster R-CNN

To enhance the performance, the original Faster R-CNN has been adapted to use ResNet50 as its backbone network instead of VGG16. In 2014, Simonyan et al. [21] first introduced the VGG16 network for large-scale image recognition in their study titled "Very Deep Convolutional Networks for Large-Scale Image Recognition". This convolutional neural network includes 13 convolutional layers, five max pooling layers, three fully connected layers, and a single softmax layer. In total, the VGG Net consists of 138 million parameters. However, this network takes a high amount of time to train on image data. It is also a rather complex model with a high computational cost. Furthermore, this model has an issue regarding the vanishing gradient problem. The architecture of a VGG16 network is shown in Fig. 4.

The ResNet model was proposed by He et al. [13], and in 2014, it was awarded in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The fundamental building element of ResNet is its residual block where each residual unit's (RU) input is combined with the unit's output and used as an input to the following RU. As a result, unlike VGG, ResNet is able to train deep neural layers without compromising performance. The ResNet50 architecture is a ResNet variant with 50

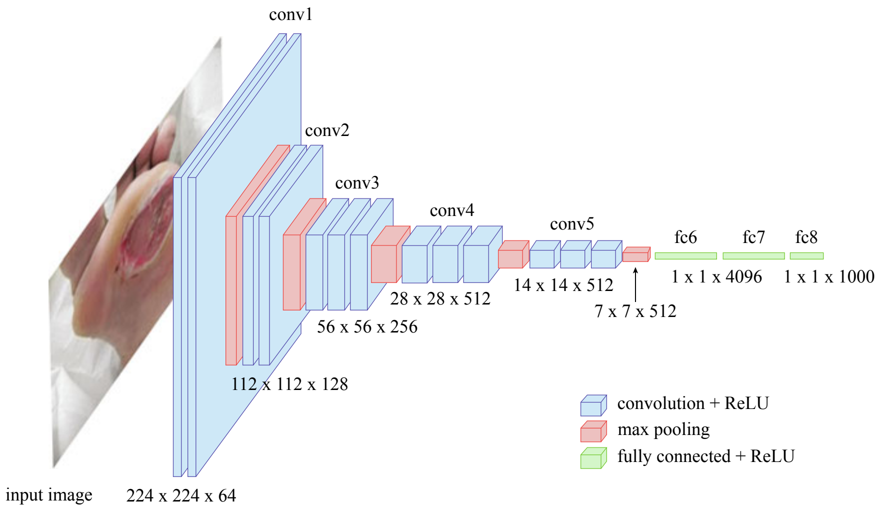


Fig. 4 Architecture of VGG16 network

deep layers. Figure 5 shows a simplified architecture of the ResNet50 model. The reason behind choosing ResNet50 in our case is because according to Ren et al. [16] it has been scientifically proved that when compared to some popular CNNs such as VGG and Inception [23, 24, 27], the pre-trained ResNet50 performs significantly better.

Now, Faster R-CNN makes use of a shared convolutional layer with the region proposal network which greatly reduces the computational cost for region proposals. In the original implementation, the value of proposed regions is set to 300, but it has been proved by Fan et al. [7] that reducing this number can lead to improved precision and response time. So, while training the model, instead of using the number 300 we changed the value to 100 ROIs. After careful analysis of the DFUC2020 dataset, we found that in terms of the size and form of the ulcers, there is a wide range of variance. Some of the lesions are too small that the standard anchors defined in the original Faster R-CNN algorithm may not detect those cases. To address this issue, we added a smaller scale of anchors to the original ones. So, in the training phase, there was a total of 12 anchors used by the region proposal network. These anchors covered a total size of 64×64, 128×128, 256×256, 512×512 along with the aspect ratios 1:1, 1:2, and 2:1. Adding this extra scale helped the algorithm to detect smaller ulcers as well as improve the detection accuracy [22].

4.4 Implementation and Training

Both the original and the enhanced version of Faster R-CNN have been implemented while conducting the experiments. The standard version of Faster R-CNN was used by Yap et al. [12] in their works. Our implementation of the methods was based on the codebase provided in https://github.com/kbardool/keras-frcnn repository. Before training the model, we prepared our dataset by splitting the 2000 training images of the DFUC 2020 dataset into two parts: 1600 images for training and 400 images for testing. As in the DFUC 2020 dataset, no ground truth has been provided for the validation and testing images, and we only used the training images and divided them into training and testing parts for our experiment. While training, Monte Carlo cross-validation was applied and the training set was randomly divided into 85% images for training and 15% images for validation. The model was trained for a total of 100 epochs with a learning rate of 1×10^{-5} for the first 70 epochs and with a learning rate of 1×10^{-6} for the remaining 30 epochs.

5 Experiment and Result

In this experiment, two different approaches were tested to detect diabetic foot ulcers from a patient's foot images. The aim was to find out the benefits of our adapted version compared to the standard version of the detector. In both cases, the models

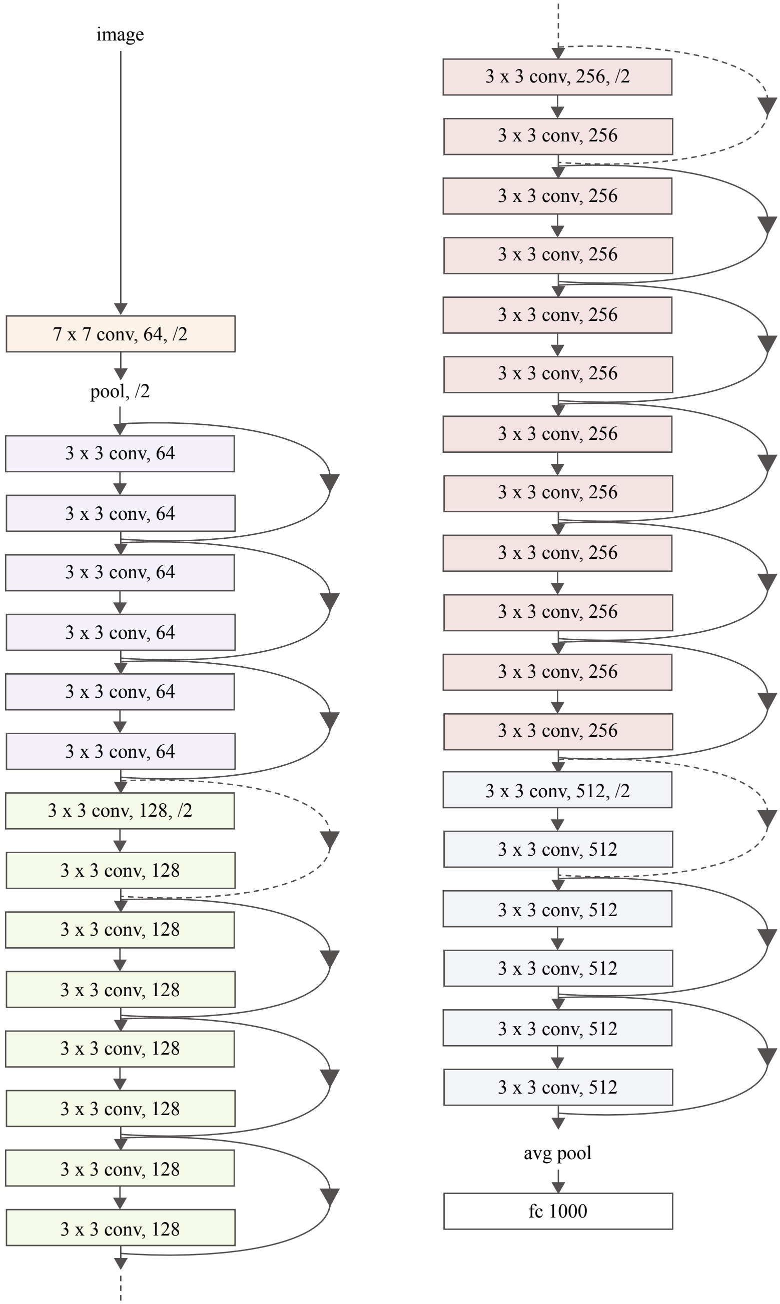


Fig. 5 Architecture of ResNet50 (only 34 layers are shown for simplicity)

Table 1 Performance of the DFU detection methods on the DFU dataset

Methods	TP	FP	Precision	Recall	F1-Score	mAP
Original faster R-CNN	240	104	0.698	0.811	0.750	0.678
Amended faster R-CNN	282	83	0.773	0.890	0.827	0.710

were trained for 100 epochs. After training, the models were applied to the 400 test images for DFU detection and comparison. The algorithms were implemented using Python 3 and TensorFlow 2.0. The experiments were conducted on a computer that has an Intel i5-4590 @3.7GHz CPU, NVIDIA GeForce GTX 1050Ti 4GB GPU, and 8GB DDR4 RAM. Table 1 summarizes the results obtained from the two methods described above. It displays the number of true positives (TP), the number of false positives (FP), precision, recall, F1-score, and mean average precision (mAP) of both detectors.

The precision, recall, F1-score, and mAP can be calculated using the following equations:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

$$\text{mAP} = \frac{1}{N} \sum_{k=1}^{k=N} \text{AP}_k \tag{4}$$

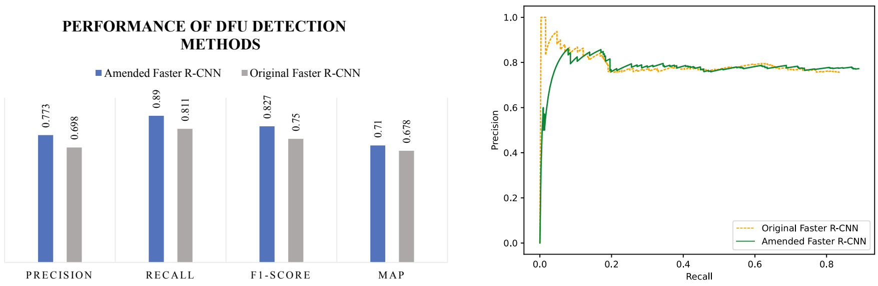


Fig. 6 Comparison of performance and precision–recall curves for different DFU detection methods

Here,

TP = total number of true positives.
FP = total number of false positives.
AP_k = average precision of class k.
N = number of classes.

From Table 1 and Fig. 6, it is clear that Faster R-CNN with pre-trained ResNet50 as its backbone CNN outperforms the pure Faster R-CNN in terms of performance. Our proposed method obtained 7.5, 7.9, 7.7, and 3.5% improvement on original Faster R-CNN in terms of precision, recall, F1-score, and mAP, respectively. Figure 7 shows some of the outputs of the proposed detection model.

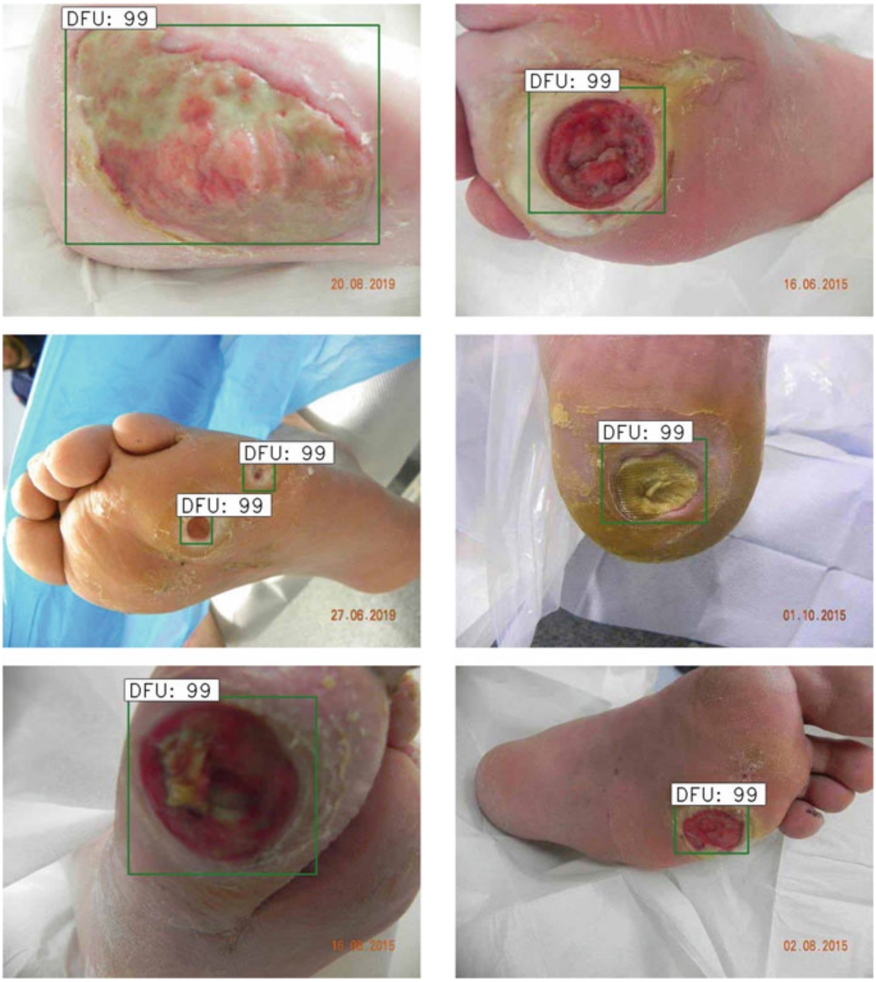


Fig. 7 DFU detection results using amended faster R-CNN

6 Conclusion

The advancement in deep learning has opened up new windows for medical image analysis by allowing the detection of textural patterns in image data. For instance, Faster R-CNN has demonstrated remarkable performance in a range of medical applications. In this paper, we presented an enhanced version of Faster R-CNN for the detection of DFU. To enhance the algorithm's performance, we changed the backbone CNN of the original Faster R-CNN from VGG16 to ResNet50. We used data augmentation and changed the number of ROIs and the anchor scales to obtain high precision and increase the detection accuracy of small lesions. We experimented with the DFUC2020 dataset which is a purposely built dataset for training and validation of automated DFU detection method. The outcomes of our experiments demonstrate that in terms of precision, recall, F1-score, and mean average precision, Faster R-CNN with some fine-tuned parameters and ResNet50 as its backbone performs better than its standard state-of-the-art implementation.

Acknowledgements We gratefully acknowledge the support of the Department of Computing and Mathematics of Manchester Metropolitan University, Manchester, UK, for providing us with the DFUC 2020 dataset.

References

1. Armstrong, D.G., Lavery, L.A., Harkless, L.B.: Validation of a diabetic wound classification system: the contribution of depth, infection, and ischemia to risk of amputation. Diabet. Care **21**(5), 855–859 (1998)
2. Brown, R., Ploderer, B., Da Seng, L.S., Lazzarini, P., Van Netten, J.: Myfootcare: a mobile self-tracking tool to promote self-care amongst people with diabetic foot ulcers. In: Proceedings of the 29th Australian Conference on Computer-Human Interaction. pp. 462–466 (2017)
3. Cassidy, B., Reeves, N., Joseph, P.: Dfuc2020: analysis towards diabetic foot ulcer detection. arxiv preprint arxiv: 2004.11853 (2020)
4. Cavanagh, P., Attinger, C., Abbas, Z., Bal, A., Rojas, N., Xu, Z.R.: Cost of treating diabetic foot ulcers in five different countries. Diabet/Metab. Res. Rev. **28**, 107–111 (2012)
5. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
6. Everingham, M., Winn, J.: The pascal visual object classes challenge 2007 (voc2007) development kit (2009)
7. Fan, Q., Brown, L., Smith, J.: A closer look at faster R-CNN for vehicle detection. In: 2016 IEEE Intelligent Vehicles Symposium (IV), pp. 124–129. IEEE (2016)
8. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
9. Goyal, M., Reeves, N.D., Davison, A.K., Rajbhandari, S., Spragg, J., Yap, M.H.: Dfunet: Convolutional neural networks for diabetic foot ulcer classification. IEEE Trans. Emerg. Topics Comput. Intell. **4**(5), 728–739 (2018)
10. Goyal, M., Reeves, N.D., Rajbhandari, S., Ahmad, N., Wang, C., Yap, M.H.: Recognition of ischaemia and infection in diabetic foot ulcers: dataset and techniques. Comput. Bio. Med. **117**, 103616 (2020)

11. Goyal, M., Reeves, N.D., Rajbhandari, S., Yap, M.H.: Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices. IEEE J. Biomed. Health Inf. **23**(4), 1730–1741 (2018)
12. Goyal, M., Yap, M.H., Reeves, N.D., Rajbhandari, S., Spragg, J.: Fully convolutional networks for diabetic foot ulcer segmentation. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 618–623. IEEE (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Liu, C., van Netten, J.J., Van Baal, J.G., Bus, S.A., van Der Heijden, F.: Automatic detection of diabetic foot complications with infrared thermography by asymmetric analysis. J. Biomed. Opt. **20**(2), 026003 (2015)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Proc. Syst. **28**, 91–99 (2015)
16. Ren, Y., Zhu, C., Xiao, S.: Object detection based on fast/faster rcnn employing fully convolutional architectures. Math. Prob. Eng. **2018** (2018)
17. Saha, S., Nassisi, M., Wang, M., Lindenberg, S., Sadda, S., Hu, Z.J., et al.: Automated detection and classification of early AMD biomarkers using deep learning. Sci. Rep. **9**(1), 1–9 (2019)
18. Saha, S.K., Fernando, B., Cuadros, J., Xiao, D., Kanagasingam, Y.: Automated quality assessment of colour fundus images for diabetic retinopathy screening in telemedicine. J. Digit Imag. **31**(6), 869–878 (2018)
19. Saha, S.K., Fernando, B., Xiao, D., Tay-Kearney, M.L., Kanagasingam, Y.: Deep learning for automatic detection and classification of microaneurysms, hard and soft exudates, and hemorrhages for diabetic retinopathy diagnosis. Invest. Ophthalmol. Vis. Sci. **57**(12), 5962–5962 (2016)
20. Saminathan, J., Sasikala, M., Narayanamurthy, V., Rajesh, K., Arvind, R.: Computer aided detection of diabetic foot ulcer using asymmetry analysis of texture and temperature features. Infrar. Phys. Technol. **105**, 103219 (2020)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Sun, X., Wu, P., Hoi, S.C.: Face detection using deep learning: an improved faster RCNN approach. Neurocomputing **299**, 42–50 (2018)
23. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
25. Yap, M.H., Cassidy, B., Pappachan, J.M., O'Shea, C., Gillespie, D., Reeves, N.: Analysis towards classification of infection and ischaemia of diabetic foot ulcers. arXiv preprint arXiv:2104.03068 (2021)
26. Yap, M.H., Hachiuma, R., Alavi, A., Brüngel, R., Cassidy, B., Goyal, M., Zhu, H., Rückert, J., Olshansky, M., Huang, X., et al.: Deep learning in diabetic foot ulcers detection: a comprehensive evaluation. Comput. Bio. Med. 104596 (2021)
27. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp. 818–833. Springer (2014)

Public Sentiment Analysis on COVID-19 Vaccination from Social Media Comments in Bangladesh

Ashfaquer Rahat Siddique, B. M. Anisuzzaman, Tareq Al Mamun, Md. Saddam Hossain Mukta, and Khondaker A. Mamun

Abstract Corona is a super virus that is storming the world, and COVID-19 has already been recognized as one of the deadliest pandemics. Herd immunity is the only solution to stop such epidemic, and a vaccine is the fastest means to reach there. But we have to remember that, it is the vaccination, not the vaccine that can stop the virus. To control the disease, government has to take proper campaign to educate and make people aware about the necessity of taking the vaccine eradicating the misinformation about it. There are different opinions and judgements present among the general people. In our study, we have used social media comments from different posts and news as data to predict users' sentiment in the context of Bangladeshi demographics. We have scrapped the comments from social media platforms and analyzed their sentiments. We have used KNN, Gaussian, and Naive Bayes classifiers to check which one gives better result. From the study, we could find out that 38.81% people show positive, 27.44% people show negative, and around 33.74% people show neutral sentiment toward the vaccination. The Bangladeshi government has to take proper steps to remove the fear of vaccine before implementing the mass vaccination program as the number of people with positive sentiment is not very high.

Keywords COVID-19 · Vaccine · Social media comments · Bangladesh

1 Introduction

COVID-19 pandemic has shaken the world. Unthinkable has happened—the world stood still, confining huge number of its population inside their homes. In March 2020, COVID-19 was declared as a pandemic by World Health Organization (WHO), and still now, it is storming the world. Though public health measures like wearing mask, maintaining social distancing, etc., are somehow effective in limiting the spread of this virus, but ultimately, an effective vaccine is needed to stop such pandemic. Because herd immunity is the final solution to pandemic and infectious diseases,

A. R. Siddique · B. M. Anisuzzaman · T. Al Mamun · Md. S. H. Mukta · K. A. Mamun (✉)
Department of Computer Science and Engineering, United International University Dhaka, Dhaka, Bangladesh
e-mail: mamun@cse.uiu.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_41

and vaccination is the most effective public health measure against such diseases [1]. Research shown that, in order to achieve some level of herd immunity, at least 60–80% of the population need to be vaccinated [2, 3]. Desperate efforts were undertaken globally to develop COVID-19 vaccine with in short period of time, and now, the vaccines are reality. But it is vaccination, not the vaccine that can prevent the pandemic [3]. In order vaccine to be effective, a large portion of the population need to come under the coverage of vaccination.

But bringing major segment of the population under vaccination coverage is not an easy task. Public support for the vaccination is of paramount importance. WHO has recognized vaccine hesitancy as one of the top ten health threats to global health [4]. Misinformation and anti-vaccination movement take people toward vaccine hesitancy. Under such situation, it is extremely important to ascertain the sentiment of the mass people toward COVID-19 vaccination which will indicate the willingness to be vaccinated. Therefore, in order to design a successful COVID-19 vaccination campaign, first, the government needs to have a clear idea about the sentiment and attitude of the people toward COVID-19 vaccine.

Our study is focused on the analysis of Bangladeshi citizens' perspectives on COVID-19 vaccines by answering one simple questions—What is the attitude of Bangladeshi citizens toward the COVID-19 vaccine?

Sentiment analysis or opinion mining task plays an important role on public opinion and national security analysis [5] which is defined as the field of study that analyzes people's opinions, evaluations, attitudes, sentiments, appraisals, and emotions toward entities such as services, products, individuals, organizations, topics, events, issues, and their attributes [6]. In this study, an attempt has been made to analyze the public sentiment of Bangladeshi people about COVID-19 vaccine from social media comments. This may help the policy makers to design a successful mass vaccination campaign. Here, social media comment has been chosen as the source of data because of the growing popularity of social media as major public opinion finder and its ability to reflect public sentiment [7].

The rest of the paper is structured as following: The related literature is discussed in Sect. 2, the methodology of the study is given in Sect. 3, data collection process is explained in Sect. 4, building sentiment analysis model is in Sect. 5, and finally, the result analysis and discussion are in Sect. 6. Section 7 contains the conclusion of the paper.

2 Related Literature

A good number of studies have been done to analyze the public sentiment and opinion about COVID-19 vaccine from social media data in different countries. Yousefinaghani et al. (2021) studied total 4,552,652 publicly available tweets that were posted between January 2020 and January 2021 and found out that the positive sentiment about the COVID-19 vaccine was the most prevalent sentiment on Twitter with higher engagement which is defined by number of retweets, likes, favorites, and replies [8]. They have also found out that the tweet accounts producing anti-vaccine contents are mostly generated by the Twitter bots creating automatic content, political activists, authors, and artists, and though, they generated a significant number of posts but with low engagement. Oppositely, pro-vaccine accounts were the renowned individuals and organizations generating less umber of posts but with higher engagement.

Puria et al. (2020) studied the vaccine content on social media and tried to identify how social media platforms are used for propagating vaccine hesitancy and explore next steps to improve health literacy and foster public trust in vaccination by using social media [9]. They suggested digital health strategies to get out of vaccine misinformation on social media by leveraging social media platforms, promoting information accuracy, targeting parents and youth, recruiting research participants and structural change to social media networks where different social media companies agreed to through joint statement to combat misinformation about the virus. At the same time, Oehler suggested that health agencies, major medical organizations, and government Websites should be improved as they have given very low priority on developing their presence in social media [10].

Piedrahita-Vald´es et al. (2021) used a hybrid approach which is a combination of supervised machine learning (support vector machine) approach and lexicon-based approach, to perform a sentiment analysis on 1,499,227 vaccine-related English and Spanish tweets published within 1st June 2011 to 30th April 2019 in languages [11]. The study was not confined to a geographical locations, and they have found out that sentiment polarity (positive, negative, or neutral) showed significant variations based on the location, and positive tweets had higher engagement than that of negative tweets. But the engagement of the tweet for vaccine (not a specific vaccine like COVID-19) varies depending on data set [11].

Pristiyono et al. (2021) studied sentiment analysis of COVID-19 vaccine in Indonesia using Na¨ıve Bayes algorithm on Twitter data with the keyword ‘COVID-19’ [12]. They collected the twitter data during the second and third weeks of January 2021. Their analysis during that period of time showed 39% positive sentiment, 56% negative sentiment, and 1% neutral sentiment. Negative sentiment about the COVID-19 vaccine was the most prevalent in that period, and the probable cause identified as the lack of trust about the safety of the vaccine.

Villavicencio et al. (2021) collected 11,974 tweets to analyze the sentiment about COVID-19 vaccine in Philippines. They have used the Na¨ıve Bayes classification algorithm with 81.77% accuracy. They have found that 83% of the tweets in the Philippines were positive, 9% were neutral, and 8% were negative about COVID vaccination [13].

Praveen et al. (2021) studied the attitude of Indian citizens toward COVID-19 vaccine [14]. They also have used the Twitter data. In their study, they have tried to find out the answer of two questions. Firstly, they tried to find out how the general perception of Indian citizens about the COVID-19 vaccine changes over different months of COVID-19 crises. Secondly, to determine the major issues that concern the general public regarding the COVID-19 vaccine. Their findings were that 47% of Twitter posts about COVID-19 vaccine were with a neutral tone, 17% were in a negative tone, and 35% had positive sentiments regarding vaccine. Fear of health and allergic reactions from the vaccine were the two most significant factors about which the Indian citizens were concerned about regarding the COVID-19 vaccine.

Observational study has been done on Facebook and Twitter comments of USA and UK general people to know about their opinion on COVID-19 vaccine [15]. Both the countries have positive sentiment as the most prevailing sentiment and their findings were broadly correlated with the findings of the nation-wide surveys in both countries. China is at the front in the world in COVID-19 pandemic. Yin et al. (2021) tried to find out the determinants of COVID-19 vaccine acceptance in China by studying Weibo messages about COVID-19 vaccines from January to October 2020 [16]. Their findings were that Chinese are more inclined toward positive side of the vaccine and also proud of China's involvement with vaccine development.

3 Methodology

The purpose of this study is to analyze the public sentiment of Bangladeshi people toward the COVID-19 vaccine from social media comments. The outcome of this study may assist policymakers in understanding the general perception about the COVID-19 vaccine and help to decide the course of action that will bring, if not all but the majority of the population under the coverage of the vaccination program.

It is clearly evident form the related literature review that most of the studies on sentiment analysis used data from Twitter. It has been identified by Sinnenberg et al. (2017) and Zhang et al. (2018) that majority of the qualitative and quantitative health related research in social media are focused on Twitter. This is mostly because of the ease with which public data can be extracted which are already in text form and the size of the data [17, 18]. But in our study, we have discarded Twitter data since it is not a popular social media platform in Bangladesh [19]. Only 2.26% of social media users use Twitter where as 88.08% and 6.53% users use Facebook and YouTube, respectively. Therefore, in our study, we have used comments from Facebook and YouTube.

The comments of Youtube and Facebook, being the primary source of data, were collected. These comments are then translated and scrutinized manually to remove irrelevant and unintelligible comments. In order to clean the data, the extracted comments were cleaned by removing background noise such as Html, hashtags, stop words, and emogies. WordNetLemmatizer of Python, PorterStemmer, and nltk.tokenize were used for data cleaning. These preprocessed data were then applied to a text analysis program, named linguistic inquiry and word count (LIWC). LIWC is a text analysis program capable of calculating the percentage of words in a text that falls into one of over 70 categories.

Then SPSS (an IBM-made statistical analysis software package) was used for statistical analysis. We have used Fishers-correlation to identify the most significant features as per their relevance and importance. Basing on that, the most essential 20 features were selected. Finally, three famous algorithms were applied to build the model. The algorithms used are K-nearest neighbor, Gaussian, and Naive Bayes algorithms. Output from these three algorithms is compared and analyzed.

4 Data Collection

Comments from social media have been used as data sources since public sentiment is vividly reflected through it. News and posts on YouTube and Facebook on the COVID-19 vaccine were explored. We searched only the news and posts which relate to Bangladesh. The comments of such news and posts were scrapped at first. More than 10,000 comments were collected. Some of these comments were in English, some were in Bengali, and some were mixed. At first, all these comments were sorted one by one, and irrelevant duplicate comments were discarded. Only the comments bearing the sentiments of general people like positive, negative, or neutral were kept. After careful scrutiny, we finally kept 1793 comments and translated them using google translate. Since google translations were not fully correct, all 1793 comments were rechecked to ensure the translation accuracy, and where necessary the comments were re-translated for proper translation. Then, the comments were labeled in three different categories as positive, negative, and neutral. It was done separately by three individuals, and the majority voting was used for selecting the final sentiment. Finally, the data set contained a total of 1793 data points.

5 Sentiment Analysis Building Models

In order to build the model, first the comments data set that contains total 1753 comments, were cleaned by removing background noise such as Html, hashtags, stop words, and emogies. WordNetLemmatize, stop words, PorterStemmer, nltk.tokenize, and RegexpTokenizer were used for cleaning and preprocessing the comments data set. At this stage, each comment is converted to an individual text file, which means that total 1793 text files were created. These text files were then passed through a text analysis application called linguistic inquiry and word count (LIWC). LIWC is a tool that analyze the text data and categorize each word into 72 predefined categories with associated score values. The results from LIWC were added to an excel files along with the label of corresponding sentiment of each comment. The excel file was analyzed by Statistical Package for the Social Sciences (SPSS), an IBM-made statistical analysis software package. The comments were analyzed, and out of 72 LIWC output features, 63 features were kept as other features did not had any value for our comments, hence were discarded. Then, Fishers-correlation was used for selecting the features as per their relatedness and importance. Twenty features were selected out of 63 for our model which were 'pronoun', 'ipron', 'affect', 'posemo', 'bio', 'health', 'verb', 'present', 'ppron', 'you', 'percept', 'hear', 'relative', 'time', 'we', 'auxverb', 'future', 'swear', 'anger', 'number'. We used three different classifiers, namely K-nearest neighbor, Gaussian, and Naive Bayes classifier so that we can compare their results. We ran our models with 20 features as mentioned above and with 10% data as testing data.

6 Results and Discussion

After the analysis, our result shows that 38.81% of Bangladeshi population possess positive sentiment about COVID-19 vaccine, whereas 27.44% bears negative sentiment. 33.74% population are neutral about the COVID vaccine (Tables 1, 2, and 3).

We analyzed the social media comment data set with three different classifiers as mentioned before. KNN, Gaussian, and Naive Bayes classifiers are used on testing data set. The classifiers' result comparison is as following.

From the above comparison, it is evident that overall, none of the classifier's accuracy is satisfactory. Out of three classifiers, Naive Bayes classifier performed comparatively better with only 46% accuracy. Different values of precision, recall, F1-score, and support for each label are given below.

For the analysis, the KNN classifier was used since it is a popularly used method in machine learning and pattern recognition for its simplicity and efficacy. The Naive

Table 1 Sentiment of social media comments

Label	Size	Percentage (%)
Positive	696	38.81
Negative	492	27.44
Neutral	605	33.74

Table 2 Accuracy of different classification models

Classifier	Accuracy
KNN	0.3833
Gaussian	0.4222
Naive Bayes	0.4611

Table 3 Performance of the classifiers

Label	Precision	Recall	F1-score	Support
Negative	0.36	0.43	0.39	47
Neutral	0.34	0.48	0.40	54
Positive	0.49	0.29	0.37	79

Bayes classifier is also a probabilistic simple classifier commonly used as the baseline in text classification. It is based on the assumption that all features are independent of each other, given the category variable. We have also used Gaussian classifier in our analysis, but a better result was achieved by Naive Bayes classifier.

Through content analysis and training/testing with different classifiers, this study pins down the fact that their exists COVID-19 vaccine reluctance among the general people of Bangladesh. Only 38.31% are positive about the vaccine and the rest either negative or neutral. This should be a concern for the government and policymakers in order to ensure a successful mass vaccination. Unless the government and the authority can convince the majority of the population about the positive results that the vaccine may bring, the objective of mass vaccination may not be achieved.

Concerns about safety, suspicions about the effectiveness and side effects, source of the vaccine, overall lack of understanding regarding the COVID-19 vaccine, and absence of messages from authoritative people were among the primary themes that emerged for people being vaccine reluctant. Bangladeshi government needs to take necessary steps to eradicate the misinformation about the vaccine before implementing the process of mass vaccination.

The main limitation of this study was the small data set which may not reflect the true picture of the general sentiment of the mass population of Bangladesh. Besides, a big number of the people do not even use social media. In our study, we did not consider the timeline with which public sentiment may change. Therefore, for future work, such study may be conducted on time period basis. For example, month wise so as to identify the change of public sentiment about COVID vaccine with respect to time. Besides, other classifiers may also be explored for better result.

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it is mentioned within a given text and the more important it is. Word clouds depicting the most commonly used word stems in COVID-19 vaccine-related posts have been created for positive, negative, and neutral sentiment which are given below.

7 Conclusion

It is needless to say that it is of paramount importance to return to the normal way of life that has been taken away by coronavirus. Only the herd immunity can return that normalcy, and the mass vaccination is the solution to reach faster herd immunity. Now, we have the vaccine in our hand, but we must remember that it the vaccination, not the vaccine that can stop this virus. Therefore, mass vaccination program is the key to stop this pandemic.

For a successful vaccination campaign, the positive sentiment among the population is the most important factor. Necessary steps are to be taken to remove the misinformation about the COVID vaccine among the people so that their negative or even neutral sentiment turns into positive one. An effective advertising campaign needs to be adopted that will motivate and encourage the general public to take

Fig. 1 Word cloud for positive sentiment

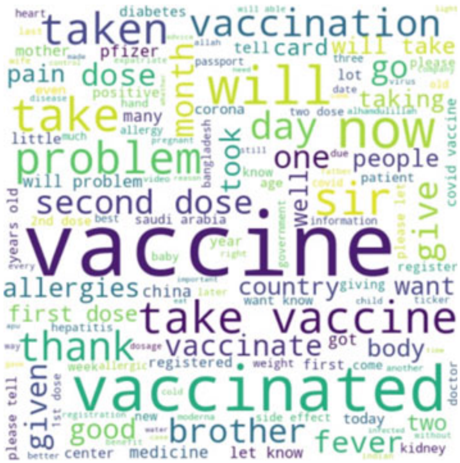
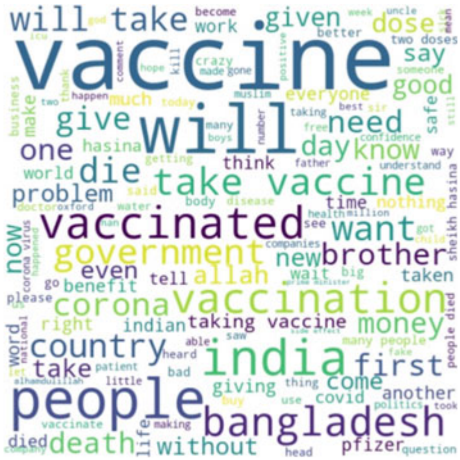


Fig. 2 Word cloud for negative sentiment



vaccines. This study may help the policymakers and government officials to understand and design a successful mass vaccination program for Bangladesh (Figs. 1, 2 and 3).

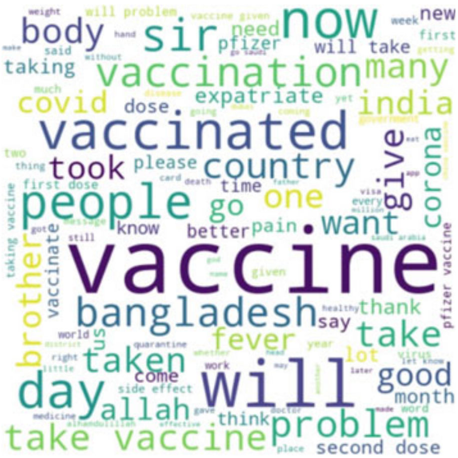


Fig. 3 Word cloud for neutral sentiment

References

1. Andre, F.E., Booy, R., Bock, H.L., Clemens, J., Datta, S.K., John, T.J., Lee, B.W., Lolekha, S., Peltola, H., Ruff, T., et al.: Vaccination greatly reduces disease, disability, death and inequity worldwide. Bull. World Health Organ. **86**, 140–146 (2008)
2. Aguas, R., Corder, R.M., King, J.G., Goncalves, G., Ferreira, M.U., Gomes, M.G.M.: Herd immunity thresholds for sars-cov-2 estimated from unfolding epidemics. medRxiv (2020)
3. Orenstein, W.A., Ahmed, R.: Simply put: vaccination saves lives (2017)
4. Qayum, I.: Top ten global health threats for 2019: the who list. J. Rehman Med. Inst. **5**(2), 01–02 (2019)
5. Yue, L., Chen, W., Li, X., Zuo, W., Yin, M.: A survey of sentiment analysis in social media. Knowl. Inf. Syst. **60**(2), 617–663 (2019)
6. Bing, L.: Sentiment analysis and opinion mining (synthesis lectures on human language technologies). University of Illinois, Chicago, IL, USA (2012)
7. Patil, H.P., Atique, M.: Sentiment analysis for social media: a survey. In: 2015 2nd International Conference on Information Science and Security (ICISS), pp. 1–4 (2015). https://doi.org/10.1109/ICISSEC.2015.7371033
8. Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., Sharif, S.: An analysis of covid-19 vaccine sentiments and opinions on twitter. Int. J. Infect. Dis. (2021)
9. Puri, N., Coomes, E.A., Haghbayan, H., Gunaratne, K.: Social media and vaccine hesitancy: new updates for the era of covid-19 and globalized infectious diseases. Hum. Vaccin. Immunother. **16**(11), 2586–2593 (2020)
10. Oehler, R.L.: On measles, vaccination, social media activism, and how to win back our role as our patients' best advocates. Clin. Infect. Dis. **70**(2), 338–340 (2020)
11. Piedrahita-Vald´es, H., Piedrahita-Castillo, D., Bermejo-Higuera, J., Guillem-Saiz, P., Bermejo-Higuera, J.R., Guillem-Saiz, J., Sicilia-Montalvo, J.A., Mach´ıo-Regidor, F.: Vaccine hesitancy on social media: sentiment analysis from June 2011 to April 2019. Vaccines **9**(1), 28 (2021)
12. Ritonga, M., Al Ihsan, M.A., Anjar, A., Rambe, F.H., et al.: Sentiment analysis of covid-19 vaccine in Indonesia using Na¨ıve Bayes algorithm. In: IOP Conference Series: Materials Science and Engineering. vol. 1088, p. 012045. IOP Publishing (2021)
13. Villavicencio, C., Macrohon, J.J., Inbaraj, X.A., Jeng, J.H., Hsieh, J.G.: Twitter sentiment analysis towards covid-19 vaccines in the Philippines using Na¨ıve Bayes. Information **12**(5), 204 (2021). https://doi.org/10.3390/info12050204

14. Praveen, S., Ittamalla, R., Deepak, G.: Analyzing the attitude of Indian citizens towards covid-19 vaccine—a text analytics study. Diabetes Metab. Syndr. Clin. Res. Rev. **15**(2), 595–599 (2021)
15. Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., Ali, A., Sheikh, A.: Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: observational study. J. Med. Internet Res. **23**(4), e26627 (2021)
16. Yin, F., Wu, Z., Xia, X., Ji, M., Wang, Y., Hu, Z.: Unfolding the determinants of covid-19 vaccine acceptance in China. J. Med. Internet Res. **23**(1), e26089 (2021)
17. Sinnenberg, L., Buttenheim, A.M., Padrez, K., Mancheno, C., Ungar, L., Merchant, R.M.: Twitter as a tool for health research: a systematic review. Am. J. Public Health **107**(1), e1–e8 (2017)
18. Zhang, A.J., Albrecht, L., Scott, S.D.: Using twitter for data collection with healthcare consumers: a scoping review. Int. J. Qual. Methods **17**(1), 1609406917750782 (2018)
19. Social media stats Bangladesh. https://gs.statcounter.com/social-mediastats/all/bangladesh. Last accessed 28 Oct 2021

Identification of Molecular Signatures and Pathways of Nasopharyngeal Carcinoma (NPC) Using Network-Based Approach

Alama Jannat Akhi, Kawsar Ahmed, Md. Ahsan Habib, Bikash Kumar Paul, Mohammad Rubbyat Akram, and Sujay Saha

Abstract Metabolism, gene regulation, and biological processes of the human body happen in molecular pathways, and any disruption in the pathways may directly lead to diseases. Molecular signatures can help get a closer look into cell biology and the mechanisms of human diseases. Therefore, the aim of the research is to discover molecular signatures and pathways which are active unusually in the NPC tissues to outline a few of the essential pathogenesis involved. Thus, genome-wide expression profiling of the NPC tissue is analyzed to identify pathways and distinguish them based on their functionality. Gene-disease association data is used to describe how molecular pathways of nasopharynx cancer are linked to different diseases. By analyzing the gene-disease associations, 392 genes are found out that are common with NPC and other diseases. This study also finds out that neuronal disorders and cancer diseases categories are closely associated with NPC. Protein–protein interaction (PPI) networks which are crucial for understanding cell physiology in both normal and disease states are formed to identify the shared protein groups of vari-

A. J. Akhi (✉) · K. Ahmed · Md. A. Habib · B. K. Paul
Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh
e-mail: alamajannatsct@gmail.com

K. Ahmed
e-mail: k.ahmed.bd@ieee.org

Md. A. Habib
e-mail: mahabib@mbstu.ac.bd

B. K. Paul
e-mail: bikash@mbstu.ac.bd

B. K. Paul
Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh

M. R. Akram
Deployment and Rollout Technology Division, Robi Axiata Ltd., Dhaka, Bangladesh
e-mail: rubbyat.akram@robi.com.bd

S. Saha
Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka, Bangladesh
e-mail: sujaysaha2906@ieee.org

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_42

ous diseases. Overall, this study identifies biomarkers (e.g., hub proteins, TFs, and miRNAs) that regulate gene expression and control important biological processes and molecular pathways of NPC with other diseases. This study can assist further studies to identify the NPC biomarkers and potential drug targets. Determination of the right targets will be helpful for a combined therapeutic approach.

Keywords Nasopharyngeal carcinoma · Pathways · Molecular signature · Protein–protein interaction network (PPI) · Reporter transcription factors · Reporter microRNAs

1 Introduction

Nasopharyngeal carcinoma (NPC), known as nasopharynx cancer, is a rare tumor of the head and neck area whose origin is the nasopharynx. NPC has a very complex aetiology that is not fully understood yet. Although NPC is rare within most populations, it is one of the leading types of cancer in some well-defined regions. In addition, increasing evidence proves that polygenes and cellular pathways are involved in the development and progression of NPC [1]. So far, the precise molecular mechanisms underlying the development of nasopharyngeal cancer remain unknown, limiting the potential for early detection and treatment of NPC. As a result, it is vital to explore the molecular mechanisms of nasopharyngeal carcinoma progression and identify new potential biomarkers to aid in early detection and curative treatment.

The 5-year survival rate of stages I and II NPC ranges from 72 to 90 percent. However, the 5-year survival rate of stages III and IV NPC are 55% and 30%, respectively, mostly due to a relatively high incidence of locoregional recurrence or metastasis [2]. Moreover, NPC has a poor prognosis because of late presentation of lesions, poor understanding of the molecular mechanisms, no suitable markers for early detection, and poor response to available therapies [3].

This study aims to identify the common biomarkers and pathways for NPC and other diseases. It might help in creating a combined cancer therapeutic approach in future.

1.1 Resources and Procedures

A multi-step analysis strategy is used in this proposed study (as shown in Fig. 1). The gene expression dataset from NCBI is statistically analyzed to identify differentially expressed genes (DEGs). The DEGs and functional enrichment analyses are considered to identify enriched pathways and biomarkers. Researches have already revealed that NPC patients can be affected by other diseases and diseases that share molecular pathways can be treated similarly. For this reason, DEGs of NPC and disease-genes associations are analyzed. A PPI network is formed by considering

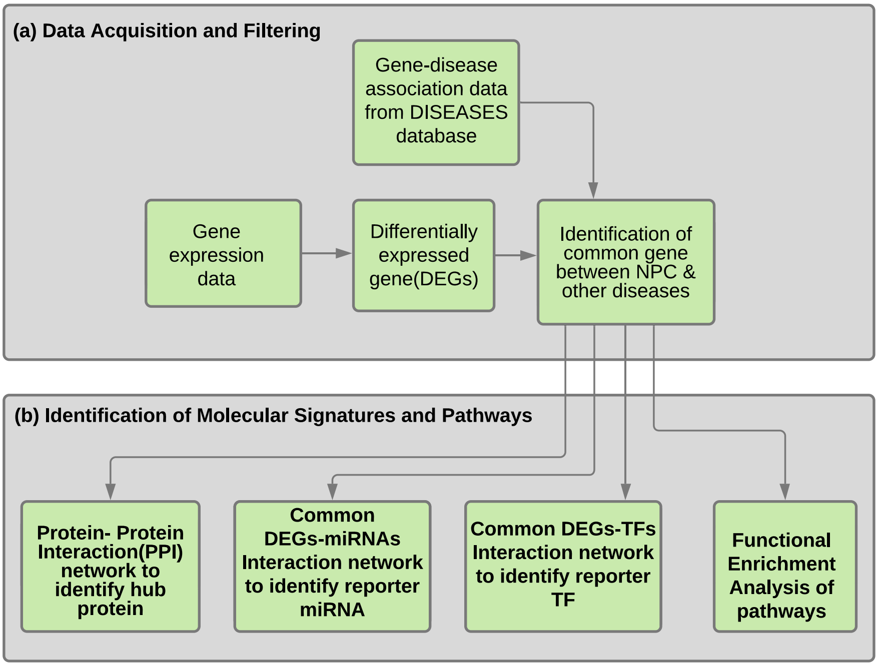


Fig. 1 Block diagram for identification of molecular signatures and pathways of NPC

the common genes that are shared by NPC and other diseases. Hub proteins are identified applying topological analysis via Network Analyst. Finally, the intermediate analysis results are combined with biomolecular networks to discover reporter biomolecules (TFs and miRNA).

1.2 *Dataset and the Statistical Methods*

Gene expression data is collected for nasopharyngeal carcinoma (GSE53819) from the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) [4]. The data sets of patients of various sexes and ages are studied. Multiple rounds of filtering, normalization, and statistical analysis are done to get DEGs. To find out genes that are differentially expressed in patients, t-test is performed. For the t-tests, an adjusted *p*-value <0.05 is chosen. The DISEASES database (https://diseases.jensenlab.org) [5] is used to gather data on gene-disease associations for this work.

The DISEASES database link human genes to diseases and the database currently hold more than 8 million associations that can be accessed via the web interface. The associated data, that is used, comes from manually curated knowledge, and the associations are causal. Curated knowledge associations are imported from the

MedlinePlus Genetics database formally known as a genetic home reference and from UniProtKB/Swiss-Prot. Causal association imply that it's almost certain that variants in these genes are associated with a disease.

The paper is organized with following sections. In Sect. 2, the relevant analysis and methodology is discussed to identify biomarkers and pathways. The results are presented in Sect. 3 with necessary Tables and Figures. Finally, the paper concludes in Sect. 4.

2 Identification of Molecular Signatures and Pathways

This section illustrates the identification process of biomarkers and pathways. It describes the network construction and analyzation to select those biomarkers. Enrichment analyzation of pathways is described at the end of this section.

2.1 *Identification of Common Genes Between NPC and Other Diseases*

DEGs of NPC and the collected gene-disease association data is used to identity the common genes between NPC and other diseases. When two diseases share at least one common dysregulated gene, they are considered to be linked [6]. Evidence is used from curated knowledge filter. Gene-disease associations aim to determine whether gene $g \in G$ is connected with disease $d \in D$, where G and D denote the specific set of human genes and human diseases, respectively. If the sets of significant down (Gy) and up (Gx) dysregulated genes associated with diseases y and x, respectively, Then, the equation

$$n = N(Gy \cap Gx)$$

represent the number of common dysregulated genes associated with both diseases y and x. The co-occurrence indicates to the number of common genes between NPC and other diseases.

Gene-disease association network (GDN) is constructed using identified common genes between NPC and other diseases using Network Analyst. A Force-Directed Graph is used in this GDN network.

2.2 *Building the PPI Network and the Analyzation of PPI*

PPI network that is essential to almost every process in a cell is constructed using the identified common 392 DEGs of NPC from the STRING database. The network is visualized and analyzed using Network Analyst, a web-based visualization software [7]. By applying topological analysis and considering the highest degree and betweenness centrality, hub proteins are identified. The KEGG pathway enrichment analyses of the PPI network are done using Network Analyst [7].

2.3 Identification of Biomolecular Features (i.e., reporter TFs and Reporter MiRNAs)

To identify Biomolecular Features (miRNAs and TFs) that regulate genes at the post-transcriptional and transcriptional level independently [8], TF target gene and miRNA-target gene interactions data are obtained from the JASPAR database and miRTarBase, respectively. These interaction networks are visualized and analyzed using Network Analyst [7].

2.4 Enrichment Analysis of the Gene Sets

Utilizing the David bioinformatics resources (https://david-d.ncifcrf.gov/) [9], pathway analysis is performed on the identified common DEGs to gain a better understanding of the molecular pathways of NPC. An adjusted p-value < 0.05 is chosen as significant to get the enrichment result.

3 Results Evaluation

The following section has shown the number of identified DEGs of NPC and the common DEGs between NPC and other diseases. The top selected hub proteins, reporter TFs, and reporter miRNAs have also shown along with the most enriched pathways.

3.1 Genomic Signature Genes

Comparing the NPC and the normal tissues, the gene expression profile of NPC patients is analyzed using microarray data from the NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgiacc=GSE53819) [10]. 18 nasopharyngeal carcinoma primary tumors and 18 non-cancerous nasopharyngeal tissues are used to perform genome-wide expressing profiling. In this analysis, 2029 differentially expressed genes (choosing p <0.05, fold change >1.0 for up regulated genes and fold change <-1.0 for down regulated genes)) are identified, where 717 genes are up regulated and 1312 genes are down regulated significantly (Fig. 2).

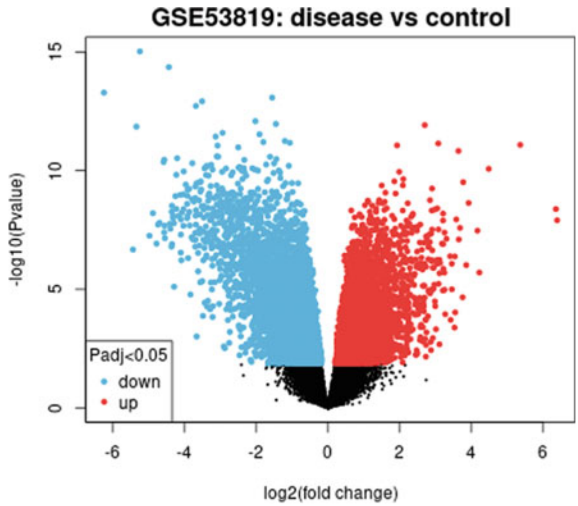


Fig. 2 vplot for up and down regulated gene

3.2 Gene-Disease Association Network Analysis

392 genes are identified as common genes between NPC and other diseases. Among those 392 genes, 160 genes are up regulated and 232 genes are down regulated. The gene-disease association network (GDN) is constructed via Network Analyst using the identified common DEGs (Fig. 3a). The list of disorders/diseases, disease genes, and associations between them are collected from the DISEASES database [5]. Using Gene Mala Cards and based on International disease classification, each dysfunction was classified into one of 19 disorder categories [11]. We found that 392 genes of NPC are associated with various diseases. The number of interconnected genes between NPC and other diseases implies that neuronal (113 genes), cancer (53 genes), respiratory (31 genes), immune system (36 genes), eye (27 genes) diseases categories are strongly associated with the NPC. Our findings point out that some genes are shared by multiple diseases. The ATM gene, for example, is shared by NPC, neuronal, and multiple conditions. As a matter of fact, neuronal and multiple diseases are linked by the ATM gene. GJA1 is a gene that is observed in NPC, cardiovascular, bone, skin, eye, and other conditions. Because of the major number of genes shared by NPC and neuronal diseases, the neuronal disease class is the most significantly related to the NPC.

3.3 Hub Proteins and Biological Markers

The PPI network is built around the common 392 DEGs in NPC and other diseases that are considered important (Fig. 3b). The topological analysis has revealed ten hub genes following: JUN, LCK, STAT1, YES1, CCND1, CDC6, ATM, CDKN1A,

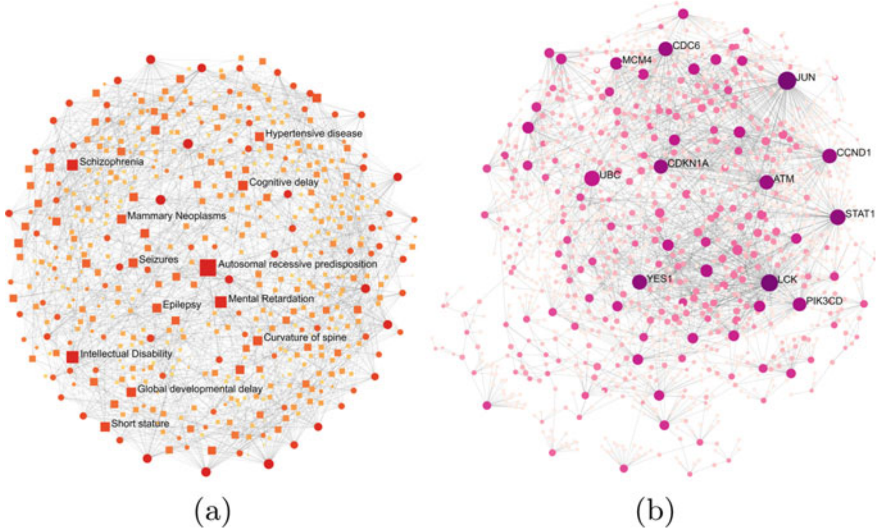


Fig. 3 **a** Gene-disease association network(GDN) using 392 common genes associated ith NPC and other diseases and **b** The PPI Network with 1543 nodes and 2575 edges using the 392 common genes associated with NPC and other diseases

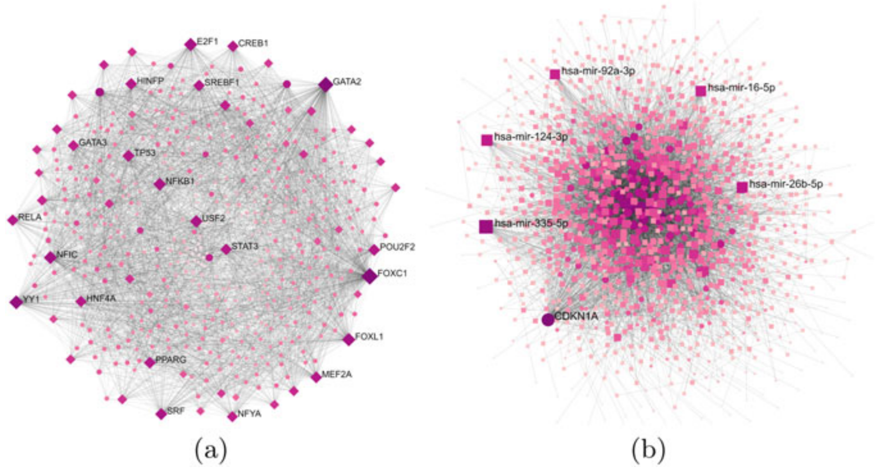


Fig. 4 **a** Interaction network between transcription factor and common 392 genes with 1176 nodes and 5289 edges and **b** the interaction network between microRNA and common 392 genes with 481 nodes and 3224 edges

PIK3CD, MCM4 (Table 1). The TFs–genes interaction network (Fig. 4a), using the topological analysis, has showed various potentially important TFs. The top ten TFs with the highest degree and betweenness centrality are as follows: FOXC1, GATA2, YY1, E2F1, NFKB1, FOXL1, NFIC, USF2, SRF, POU2F2 (Table 1). The

Table 1 Summary of molecular signatures (hub proteins, TFs, and miRNAs) in NPC

Symbol	Degree	Betweenness	Features
JUN	129	241,506.57	Hub protein
LCK	108	163,147.48	Hub protein
STAT1	84	151,583.26	Hub protein
YES1	74	74,761.12	Hub protein
CCND1	70	88,323.98	Hub protein
CDC6	68	55,018.35	Hub protein
ATM	67	82,840.54	Hub protein
CDKN1A	65	76,153.75	Hub protein
PIK3CD	59	86,344.38	Hub protein
MCM4	46	32,136.73	Hub protein
FOXC1	219	22,227.34	Reporter TF
GATA2	181	15,701.46	Reporter TF
YY1	119	5577.33	Reporter TF
FOXL1	103	4869.87	Reporter TF
E2F1	102	5746.86	Reporter TF
NFIC	96	5155.19	Reporter TF
NFKB1	88	3751.72	Reporter TF
USF2	83	3056.49	Reporter TF
SRF	76	2499.53	Reporter TF
POU2F2	74	2240.51	Reporter TF
hsa-mir-335-5p	73	50,592.34	Reporter miRNA
hsa-mir-124-3p	43	23,217.9	Reporter miRNA
hsa-mir-26b-5p	41	16,255.8	Reporter miRNA
hsa-mir-16-5p	36	13,732.51	Reporter miRNA
hsa-mir-92a-3p	32	14,765.6	Reporter miRNA
hsa-mir-192-5p	27	6753.24	Reporter miRNA
hsa-mir-1-3p	26	8263.02	Reporter miRNA
hsa-mir-215-5p	25	5953.81	Reporter miRNA
hsa-mir-6499-3p	25	5921.13	Reporter miRNA
hsa-mir-17-5p	24	7069.62	Reporter miRNA

miRNA–genes interaction network (Fig. 4b) using the topological analysis is built. By considering the highest degree and betweenness, the top ten miRNAs are chosen, and they are as follows: hsa-mir-335-5p, hsa-mir-124-3p, hsa-mir-26b-5p, hsa-mir-16-5p, hsa-mir-92a-3p, hsa-mir-192-5p, hsa-mir-1-3p, hsa-mir-215-5p, hsa-mir-6499-3p, hsa-mir-17-5p (Table 1).

3.4 Enrichment Analysation of Pathways

Ten significant KEGG pathways (Table 2) are selected using identified 392 NPC genes that are also related to other diseases.

The pathway enrichment analysis of TF–genes interaction network has shown ten significant enriched pathways including: Pathways in cancer, Proteoglycans in cancer, Focal adhesion, Breast cancer, Hepatitis B, ErbB signaling pathway, PI3K-Akt signaling pathway, AGE-RAGE signaling pathway in diabetic complications, Cell cycle, Wnt signaling pathway (Table 3).

Pathway enrichment analysis of the miRNA–genes interaction network has highlighted the following pathways: ECM–receptor interaction, AGE-RAGE signaling pathway in diabetic complications, pathways in cancer, protein digestion and absorption, focal adhesion, small cell lung cancer, amoebiasis, transcriptional misregulation in cancer, osteoclast differentiation, primary immunodeficiency (Table 4).

3.5 Discussions

In this study, interactions of NPC with other diseases are analyzed and visualized focusing on the associations of genomics, and molecular signaling pathways. These identified common genes may be candidate disease biomarkers and may also serve as potential therapeutic targets. Significant genes that have the potentiality to verify risk factors for NPC are also identified.

These associations are discovered based on the number of genes they have in common. In the category of cancer, it is found that three genes (XRCC2, ATM, BRIP1) are common with breast cancer, five genes (SRD5A2, MSMB, EPHB2, HOXB13, MSR1) are common with prostate cancer, four genes (DLEC1, RET, DDR2, ALK) are common with lungs cancer, and three genes (ATM, CDKN1A, FAT1) are common with urinary bladder cancer.

In the category of neuronal disorder, it is found that three genes (CLU, RALGPS2, UNC5C) associated with Alzheimer's disease, five genes (LAMB1, TMTC3, B3GALNT2, CEP85L, APC2) with Lissencephaly and five genes (C1orf194, GJB1, SCO2, HSPB8, PDXK) with Charcot–Marie–Tooth disease.

The present study found 392 common DEGs from the analysis. Based on the combined analysis of genomic and biomolecular network analysis, critical biomolecules are revealed (hub proteins, TFs, and miRNAs) that may indicate an association between NPC and and other diseases.

One common pathway identified in this study is PI3K-Akt signaling pathway (as given in Tables 2 and 3). One study found that PI3K-Akt signaling pathway has a significant impact on lung cancer [12] and thyroid cancer is actively connected with it [13]. From some previous studies, it was found that dysregulation of this pathway was identified in various cancers including colorectal cancer, breast cancer, and hematologic malignancies [14, 15]. Another common pathways identified are ECM-

Table 2 Pathways involved with the common genes between NPC and other diseases

KEGG IDs	Pathways	Genes involved	Fold enrichment	*p*-value
hsa04512	ECM–receptor interaction	TNXB, LAMB3, LAMB2, LAMA1, COL11A1, FN1, LAMB1, LAMC2, GP1BA, THBS2, COL1A1, COL5A1, COL4A1, COL4A3, COL4A5, SDC1, ITGA7, AGRN	6.875562219	5.05E−10
hsa05200	Pathways in cancer	RET, CDKN1A, EPAS1, LAMA1, PIK3CD, LAMC2, EDNRB, MECOM, CCND1, PLCG2, WNT4, RUNX1T1, APC2, JUN, LAMB3, JUP, FZD4, LAMB2, STAT1, ZBTB16, FZD6, WNT5A, FN1, LAMB1, TGFBR2, RUNX1, VEGFA, MAPK10, KITLG, COL4A1, COL4A3, COL4A5, FGF12	2.790463547	1.64E−07
hsa04510	Focal adhesion	JUN, TNXB, LAMB3, LAMB2, LAMA1, COL11A1, FN1, PIK3CD, LAMB1, LAMC2, THBS2, VEGFA, MAPK10, COL1A1, COL5A1, CCND1, COL4A1, COL4A3, COL4A5, ITGA7	3.22639651	1.13E−05
hsa05146	Amoebiasis	LAMB3, LAMB2, LAMA1, COL11A1, FN1, PIK3CD, LAMB1, LAMC2, COL1A1, IFNG, COL5A1, COL4A1, COL4A3, COL4A5	4.389116762	1.46E−05
hsa04974	Protein digestion and absorption	COL17A1, COL1A1, COL5A1, COL4A1, COL11A1, COL7A1, COL4A3, COL12A1, COL10A1, COL4A5, ATP1A2, KCNN4	4.531620553	5.64E−05
hsa04151	PI3K-Akt signaling pathway	CDKN1A, TNXB, LAMA1, COL11A1, PIK3CD, LAMC2, THBS2, TCL1A, CCND1, CD19, MYB, LAMB3, LAMB2, FN1, LAMB1, VEGFA, COL1A1, KITLG, COL5A1, COL4A1, COL4A3, COL4A5, ITGA7, TEK, FGF12	2.40810754	8.99E−05
hsa05202	Transcriptional misregulation in cancer	CDKN1A, MEF2C, JUP, ZBTB16, LMO2, HMGA2, PAX5, FLI1, TGFBR2, RUNX1, LYL1, TSPAN7, ATM, ERG, PROM1, RUNX1T1	3.183893083	1.32E−04
hsa05222	Small cell lung cancer	CCND1, LAMB3, COL4A1, LAMB2, LAMA1, COL4A3, FN1, COL4A5, PIK3CD, LAMB1, LAMC2	4.30059676	2.06E−04
hsa05340	Primary immunodeficiency	ZAP70, CD40LG, TNFRSF13B, LCK, CD19, BTK, TNFRSF13C	6.841858483	4.47E−04
hsa04380	Osteoclast differentiation	JUN, NCF1, STAT1, PIK3CD, TREM2, TGFBR2, MAPK10, IL1A, IFNG, LCK, BTK, PLCG2, TNFSF11	3.297820555	5.23E−04

Table 3 10 Top-ranked KEGG pathways involved in TFs–genes interaction network

Pathway	Total	Expected	Hits	*p*-value
Pathways in cancer	530	82.3	285	1.15E−101
Proteoglycans in cancer	201	31.2	140	1.04E−68
Focal adhesion	199	30.9	125	2.12E−53
Breast cancer	147	22.8	103	7.64E−51
Hepatitis B	163	25.3	108	1.60E−49
ErbB signaling pathway	85	13.2	74	7.35E−49
PI3K-Akt signaling pathway	354	55	168	4.14E−48
AGE-RAGE signaling pathway in diabetic complications	100	15.5	78	2.11E−44
Cell cycle	124	19.2	88	3.22E−44
Wnt signaling pathway	158	24.5	100	3.70E−43

Table 4 10 Top-ranked KEGG pathways involved in the miRNAs–genes interaction network

Pathway	Total	Expected	Hits	*p*-value
ECM–receptor interaction	82	2.16	16	2.73E−10
AGE-RAGE signaling pathway in diabetic complications	100	2.64	14	3.01E−07
Pathways in cancer	530	14	34	9.44E−07
Protein digestion and absorption	90	2.37	12	3.68E−06
Focal adhesion	199	5.25	18	4.55E−06
Small cell lung cancer	93	2.45	12	5.22E−06
Amoebiasis	96	2.53	12	7.30E−06
Transcriptional misregulation in cancer	186	4.9	17	7.36E−06
Osteoclast differentiation	128	3.37	13	2.99E−05
Primary immunodeficiency	37	0.976	7	4.20E−05

receptor interactions (as given in Tables 2 and 4). According to previous studies, ECM–receptor interactions are involved in cell adhesion [16].

Constructing and analyzing the PPI network are vital to understanding cell physiology in various states like normal and disease states. The hub protein ATM identified in this study shows, with the recent explosion in genomic data, ATM alterations have been revealed both in the germline as a predisposing factor for cancer and as somatic changes in tumors themselves [17]. An analysis of clinical samples showed that YES1 gene amplification existed not only in esophageal cancer but also in lung, head and neck, bladder, and other cancers, indicating that YES1 would be an attractive target for a cancer drug [18].

Transcription factors (TFs) and microRNAs are responsible for controlling important biological processes and regulating gene expression independently. Therefore, changes in these molecules have a great potentiality to provide crucial information on the dysregulation of gene expression. One of the identified TFs of this study FOXC1 plays a critical role in tumor development and metastasis. Clinical studies have demonstrated that elevated FOXC1 expression is associated with poor prognosis in many cancer subtypes [19]. YY1 has been shown to be overexpressed in many cancers including lung cancer, breast cancer, ovarian cancer, colon cancer, prostate cancer, brain cancer, cervical cancer, osteosarcoma, gastric cancer, acute myeloid leukemia, B-cell, and follicular lymphoma [20, 21]. POU2F2 is highly expressed in lung cancer cells and the involved in lung cancer progression [22]. GATA2 is closely associated with NPC [23].

Among the miRNAs identified from this study, mir-335-5p and hsa-mir-124-3p have been found associated with NPC [24]. miR-192-5p contributes to targeting SEMA3A in hepatocellular carcinoma cell, and this may be used as a target in targeted therapy and a marker for cancer behavior and prognosis [25].

4 Conclusions

The present study provides possible novel links of association between NPC and other diseases to assist the early diagnosis. This study has revealed potential biomarkers to help researchers better understand the pathogenic features and pathways of NPC. It may lead to ultimate advancement for a combined therapeutic approach and potential drug targets discoveries. The identified biomolecules presented in this study deserve experimental studies to clarify the biological roles in an association of NPC and other diseases.

References

1. Tulalamba, W., Janvilisri, T.: Nasopharyngeal carcinoma signaling pathway: an update on molecular biomarkers. Int. J. Cell Biol. **2012** (2012)
2. Chang, J., Ko, J.-Y., Hong, R.-L.: Recent advances in the treatment of nasopharyngeal carcinoma. J. Formosan Med. Assoc. Taiwan yi zhi, **103**(7), 496–510 (2004)
3. Wang, H.-Y., Zhang, M., He, P.-C., Yang, B.-J., Shao, L.-Y., Shao, W.-B.: Changes of gene expression profile in human myeloma cell line induced by thalidomide. Zhongguo shi yan xue ye xue za zhi **18**(2), 396–402 (2010)
4. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R.: Ncbi geo: mining tens of millions of expression profiles-database and tools update. Nucleic. Acids. Res. **35**(suppl_1), D760–D765 (2007)
5. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., Jensen, L.J.: Diseases: Text mining and data integration of disease-gene associations. Methods **74**, 83–89 (2015)
6. Moni, M.A., Liò, P.: Comor: a software for disease comorbidity risk assessment. J. Clin. Bioinformat. **4**(1), 1–11 (2014)
7. Zhou, G., Soufan, O., Ewald, J., Hancock, R.E., Basu, N., Xia, J.: Networkanalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucl. Acids Res. **47**(W1), W234–W241 (2019)

8. Gupta, A., Sugadev, R., Sharma, Y., Kumar, B., Khurana, P.: Mirffldb-an integrated resource for human mirna-transcription factor-target gene coregulatory networks and associated feed forward loops. bioRxiv (2020)
9. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using David bioinformatics resources. Nature Protocols **4**(1), 44–57 (2009)
10. Bao, Y., Cao, X., Luo, D., Sun, R., Peng, L., Wang, L., Yan, Y., Zheng, L., Xie, P., Cao, Y., et al.: Urokinase-type plasminogen activator receptor signaling is critical in nasopharyngeal carcinoma cell growth and metastasis. Cell Cycle **13**(12), 1958–1969 (2014)
11. Espe, S.: Malacards: the human disease database. J. Med. Library Assoc.: JMLA **106**(1), 140 (2018)
12. Shtivelman, E., Hensing, T., Simon, G.R., Dennis, P.A., Otterson, G.A., Bueno, R., Salgia, R.: Molecular pathways and therapeutic targets in lung cancer. Oncotarget **5**(6), 1392 (2014)
13. He, W., Qi, B., Zhou, Q., Lu, C., Huang, Q., Xian, L., Chen, M.: Key genes and pathways in thyroid cancer based on gene set enrichment analysis. Oncology Rep. **30**(3), 1391–1397 (2013)
14. Rahman, M.R., Islam, T., Shahjaman, M. Quinn,J.M., Holsinger, R.D., Moni, M.A.: Identification of common molecular biomarker signatures in blood and brain of Alzheimer's disease. BioRxiv 482828 (2019)
15. Yang, J., Nie, J., Ma, X., Wei, Y., Peng, Y., Wei, X.: Targeting pi3k in cancer: mechanisms and advances in clinical trials. Molecular Cancer **18**(1), 1–28 (2019)
16. Albelda, S.M., Buck, C.A.: Integrins and other cell adhesion molecules. The FASEB J. **4**(11), 2868–2880 (1990)
17. Cremona, C., Behrens, A.: Atm signalling and cancer. Oncogene **33**(26), 3351–3360 (2014)
18. Hamanaka, N., Nakanishi, Y., Mizuno, T., Horiguchi-Takei, K., Akiyama, N., Tanimura, H., Hasegawa, M., Satoh, Y., Tachibana, Y., Fujii, T., et al.: Yes1 is a targetable oncogene in cancers harboring yes1 gene amplification. Cancer Res. **79**(22), 5734–5745 (2019)
19. Han, B., Bhowmick, N., Qu, Y., Chung, S., Giuliano, A.E., Cui, X.: Foxc1: an emerging marker and therapeutic target for cancer. Oncogene **36**(28), 3957–3963 (2017)
20. Zhang, Q., Stovall, D.B., Inoue, K., Sui, G.: The oncogenic role of yin yang 1. Crit. Rev.^{TM} in Oncogenes **16**(3-4) (2011)
21. Atchison, M.L., Basu, A., Zaprazna, K., Papasani, M.: Mechanisms of yin yang 1 in oncogenesis: the importance of indirect effects. Crit. Rev^{TM} Oncogenes **16**(3-4) (2011)
22. Luo, R., Zhuo, Y., Du, Q., Xiao, R.: Pou2f2 promotes the proliferation and motility of lung cancer cells by activating ago1. BMC Pulmonary Med. **21**(1), 1–12 (2021)
23. Ye, Z., Wang, F., Yan, F., Wang, L., Li, B., Liu, T., Hu, F., Jiang, M., Li, W., Fu, Z.: Bioinformatic identification of candidate biomarkers and related transcription factors in nasopharyngeal carcinoma. World J. Surg. Oncol. **17**(1), 1–10 (2019)
24. Ye, J., Liao, Q., Zeng, X., Liu, C., Ding, Y., Liu, X., Zeng, L., Guan, T., Yuan, Y.: Microrna-124-3p inhibited progression of nasopharyngeal carcinoma by interaction with pcdh8 and the inactivation of pi3k/akt/mtor pathway. J. Cancer **12**(16), 4933–4944 (2021)
25. Yan-Chun, L., Hong-Mei, Y., Zhi-Hong, C., Qing, H., Yan-Hong, Z., Ji-Fang, W.: Microrna-192-5p promote the proliferation and metastasis of hepatocellular carcinoma cell by targeting sema3a. Appl. Immunohistochem. Molecular Morphol. **25**(4), 251–260 (2017)

IoT-Based Smart Health Monitoring System: Design, Development, and Implementation

Abdullah Al Mamun, Md. Nahidul Alam, Zahid Hasan, and Adittya Narayan Das

Abstract Health monitoring systems in hospitals, clinics, and many other health centers have experienced a significant amount of growth in the recent times. In this current pandemic situation, a lot of people are suffering from COVID-19, lungs disease, chronic disease, and various kinds of common flu, etc. The health of COVID and other common flu affected people as well as elderly people needs continuous and regular monitoring in order to avoid any disastrous situation. It is quite difficult task for health professionals or the patients to be present in person for continuous health monitoring. In this paper, we designed and implemented an IoT-based smart health monitoring system for patients, especially, the elderly, COVID-affected people and patients with chronic diseases. This healthcare monitoring system monitors the body temperature, blood oxygen levels, heart rate, and electrocardiogram (ECG) and uploads the real-time data to an open source Mosquitto (MQTT) server via Wi-Fi or a GSM modem for remote monitoring which can be accessed via a website or a mobile application. This low-cost and efficient device will help both the patient and the health professional in monitoring and diagnose of various kinds of health problems. An emergency alert system has been implemented into the system which will send text message alert to the health expert and the patient itself in case of abnormality in health data. Furthermore, health data analyzing system will help the medical professional in determining the critical patients who need special attention.

Keywords Internet of Things · Health monitoring · ECG · Temperature

1 Introduction

In Bangladesh, both in rural and urban areas, people are always at high risk of contracting various kinds of flu and viruses. In general sense, patients generally consult a doctor for medical checkup where various body health parameters are determined and monitored in order to diagnose the disease. But due to current ongoing

A. A. Mamun (✉) · Md. N. Alam · Z. Hasan · A. N. Das
Bangladesh Army International University of Science and Technology, Cumilla Cantonment, Cumilla, Bangladesh
e-mail: mamun@cer.uiu.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_43

pandemic, over-crowded hospitals, shortage of medical staff, poor economic structure, and inadequate transport facilities, it has become quite a tough task for patients and doctors for conducting regular medical checkup or continuous health monitoring. Moreover, people living in the rural areas have less access to the medical needs and necessities.

Agreeing to the most recent WHO information distributed in 2018 Coronary Heart Infection deaths in Bangladesh come to 118,287 or 15.23% of total deaths. The age-adjusted Passing Rate is 109.32 per 100,000 of population which positions Bangladesh at 115 within the world [1]. Besides, Lung Illness related deaths in Bangladesh are close to 64,762 or 8.34% of total deaths occured in 2021 [2]. The main reason behind this high rate of death is daily usage of tobacco, alcohol consumption, over stress, no physical activity, etc. Any person who is affected by chronic diseases and lung illness must have to lead his life properly at most care and should be monitored by a doctor continuously. People who has contracted COVID or dengue also need their body temperature and oxygen levels in blood to be monitored continuously. The elderly people needs their heart rate and heart condition to be monitored regularly. The most common parameters that needs to be determined before diagnosing the sickness are heart rate, body temperature, amount of oxygen in blood (%SpO2), electrocardiogram (ECG), etc. The approximate heart rate for adult human is 70–85 beats per minute [3]. The body temperature for a typical adult is around 98.6 °F or 37 °C [4]. In case of blood oxygen levels, the normal oxygen (SpO2) level is around 95–100% [5].

It is very difficult for hospital staff to monitor these important parameters for many patient's health simultaneously. During an emergency, the situation might get worse, where monitoring the critical patient 24/7 is must to reduce life threatening risk. That is why this paper indicates a monitoring system where doctor doesn't need individual thermometer, oximeter ECG machine, etc., to monitor a patient's health parameters. It is possible to easily monitor many patients at a time by using the proposed method where various body parameter sensing devices are integrated with Internet of Things (IoT). This monitoring system includes an emergency alert system where if any of the parameters of any patient reaches the critical rate determined by the health professionals, an alert text message will be sent to the patient and the corresponding doctor. The health data of patients with chronic disorders such as respiratory diseases cardiovascular diseases, diabetes or patients affected with dengue, COVID, etc., can be stored into the database which can be further examined in order to determine the critical patients. It is a very fast, accurate, and a low-cost system which will aid both the doctors and the patients.

2 Related Works

We analyzed the research objectives and outcomes as well as the drawbacks and limitations of the researches done previously. The summary of the review is given in Table 1.

Table 1 Limitations and drawbacks of previous related works

References #	Proposed	Limitations
Akash et al. [6]	IoT-Based Real-Time Health Monitoring System	In this system, various sensors were merged, and the output is showed on an OLED display. The drawback of the system is that the health parameter values have high output error%. Some of which exceeds the permissible value by the experts. Moreover, the system does not include data storage system and emergency text alert scheme
Islam et al. [7]	Development of Smart Healthcare Monitoring System in IoT Environment	The proposed system includes many sensors to determine the health but only body temperature and heart rate are the only parameters useful for real-time health care assessment. The developed system focuses more on the environmental aspect rather than health parameters. It does not include storage or emergency alert system
Ruman et al. [8]	IoT-Based Emergency Health Monitoring System	The proposed health monitoring system uses Arduino and sensors to determine health parameters. Data accuracy, limited usage of sensors and cost is the main drawback of the device. Also, it lacks emergency alert system in case of emergency
Reddy et al. [9]	Health Monitoring System Based on IoT	This system also lacks some of the important features such as data health storage and assessment mechanism, ECG monitoring, etc.
Hidayah et al. [10]	Wireless Smart Health Monitoring System Via Mobile Phone	Here, a wireless health monitoring system was developed focusing on mobile app. This paper has introduced an emergency notification mechanism for patients but limited to only body temperature and heart rate sensor
Yeri et al. [11]	IoT-Based Real-Time Health Monitoring	A complete system was developed, includes website and mobile app for monitoring and emergency text alert system. Fails to incorporate health data storage system for future and further assessment of patient's health

(continued)

Table 1 (continued)

References #	Proposed	Limitations
Valsalan et al. [12]	IoT-Based Health Monitoring System	This study includes an algorithm for determining critical patients and sends emergency notification. The limitation of this study is that it not only lacks important health parameters such as Blood oxygen level (SpO2) and ECG but also fails to present a web application for real-time monitoring as well as a proper database system for future analysis

In our proposed system, we have designed a health monitoring device that is able to mitigate most of the limitations and drawbacks shown in previous researches.

3 Proposed Design

Monitoring of vital health parameters of the elderly people, COVID, dengue, and chronic disease affected people is very important. In traditional system, the healthcare professionals collect vital data from the patients manually. This smart system includes hardware to collect vital health data from the body of a patient. The software is used to process the collected data and publish it for remote monitoring and diagnosis. The main components of the system are described below.

3.1 Main Components of the Proposed System

Main Processing Unit. The main processing unit (MPU) used in our system is ESP32 which is an Xtensa Dual Core 32-bit processor, can be seen in Fig. 1a. It is a very popular IoT device. It is low-cost, high-performance, low-power microcontroller with built-in Wi-Fi, Bluetooth. Arduino Integrated Development Environment (IDE) is used to program this device [9, 13].

ECG Sensor. The ECG sensor model used in this proposed system is AD8232 which can be seen in Fig. 1b. It is a low-cost module that measures the electrical activities generated by the heart. It collects the data of electrocardiogram from the heart of the patient using a set of electrodes. The AD8232 module includes an op amp to aid in the easy acquisition of a clean signal from the PR and QT Intervals [14].

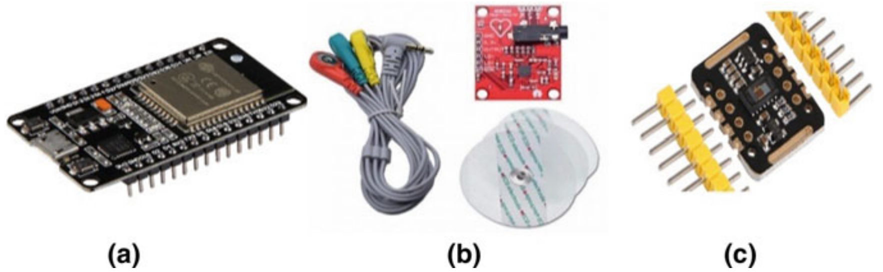


Fig. 1 ESP32 as main processor (a), ECG sensor (AD8232) (b), and MAX30102 Pulse Oximeter Sensor (c)

Pulse Oximeter Sensor. This sensor is used to measure the oxygen blood levels or SpO2 as well as the heart rate of the patient. The pulse oximeter model that we have used in our system is MAX30102 (Fig. 1c). The way this device operates is it analyzes absorbed or reflected light from red oxygenated hemoglobin or blue deoxygenated hemoglobin in the pulmonary circulation to determine a patient's blood oxygen level [15, 16].

Body Temperature Sensor. One of the most common ways to determine if a person is sick or not is by observing the body temperature. The body temperature sensor we have used in our system can be seen in Fig. 2a, which is DS18B20 waterproof sensor. This 1-wire temperature monitoring sensor is very useful to determine the temperature of a certain thing from a distance. The patients can use this device to measure the body temperature by putting the sensing probe under the arm [17].

GSM/GPRS Module. In case of no Wi-Fi service, GSM module is used for alternative way to connect to the Internet. Another important purpose of using the GSM module is to send emergency alert to the doctor and the patient in case the health data reaches a threshold pre-determined by the health professional. In our system, we have used SIM800l GSM/GPRS module (Fig. 2b) for the above-mentioned purposes. It is a miniature cellular module capable of sending and receiving text messages. Quad band frequency support makes this module capable of long-range connections [18].

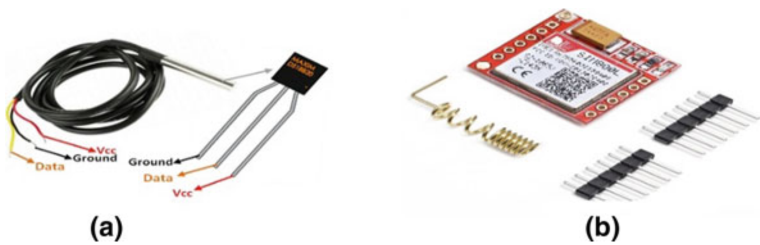


Fig. 2 DS18B20 temperature sensor (a) and SIM800L GSM module (b)

UBIDOTS. Ubidots is an IoT Application Enablement Platform (AEP) enabling System Integrators (SIs) to rapidly assemble and launch IoT applications. This is an open source IoT platform where we upload the collected data from the sensors. Anyone from anywhere in the world can access to the uploaded data with proper access. In our case, the healthcare professionals can monitor the health condition of patients using this web application [19].

3.2 Architecture and Implementation of the System

Figure 3 shows the complete architecture of the system.

The working principle of the proposed system is shown by the flow diagram as shown in Fig. 4.

As seen from the figure above, we can see that at the beginning of the system, the device will try to connect to the network via Wi-Fi or GSM module. After the connection to the Internet is successful, connection to the main server is created where all the health data will be uploaded in real time and saved in the storage for real-time monitoring. After establishing the connection, all the sensors are initialized and after completion of successful sensor initialization, vital health parameters are measured using the sensors. The health data are processed in the MPU and data are published to the MQTT cloud server for real-time monitoring. Heart beat rate, body temperature, and SpO2 levels are measured and compared with threshold values in order to send emergency alert message to the doctor and the patient. The published health data are stored in the cloud server which can be accessed later for further analyzation of the health of a patient.

Simulation and Implementation. Initially, the system was designed and simulated using Proteus Design Suit before implementing on the breadboard. During the simulation, Arduino UNO was used as the main processor as seen in Fig. 5a. After initial

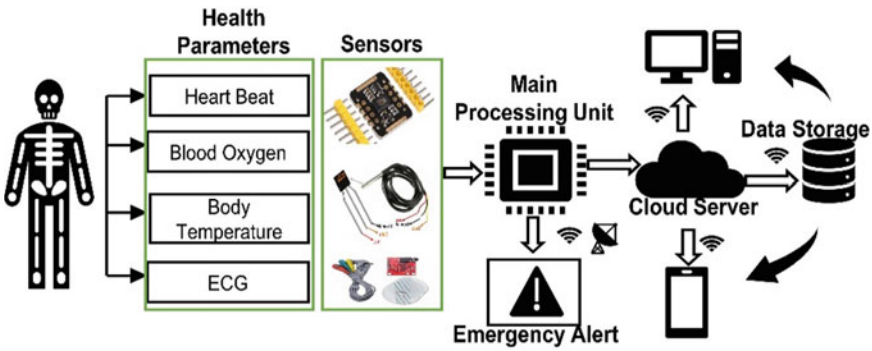


Fig. 3 Architecture of the smart health monitoring system

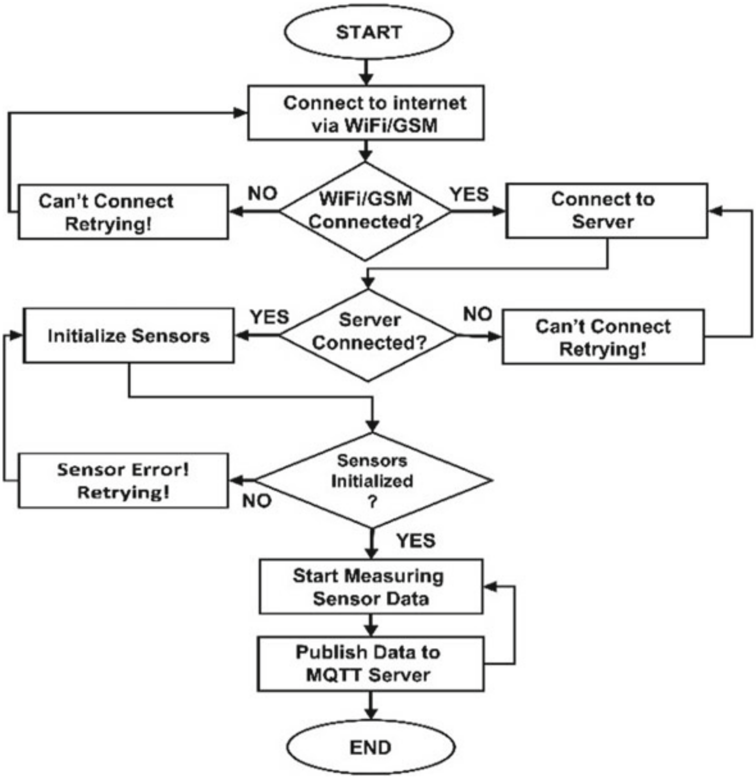


Fig. 4 Flow diagram of smart health monitoring system

simulation, the system was developed and tested on a breadboard by connecting all the sensors to the power supply system (Fig. 5b).

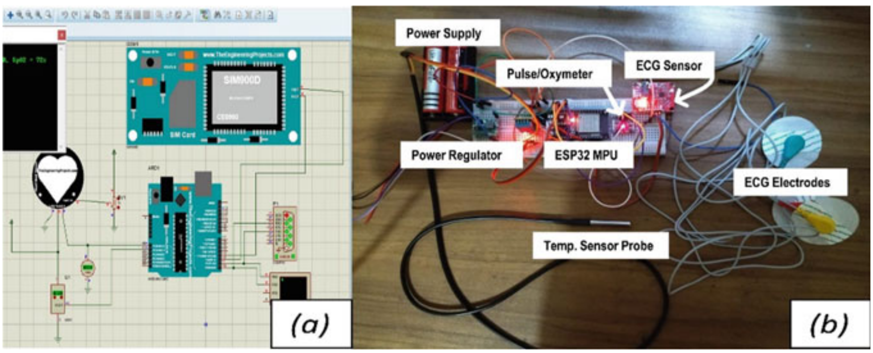


Fig. 5 Simulation (a) and implementation (b) of the system

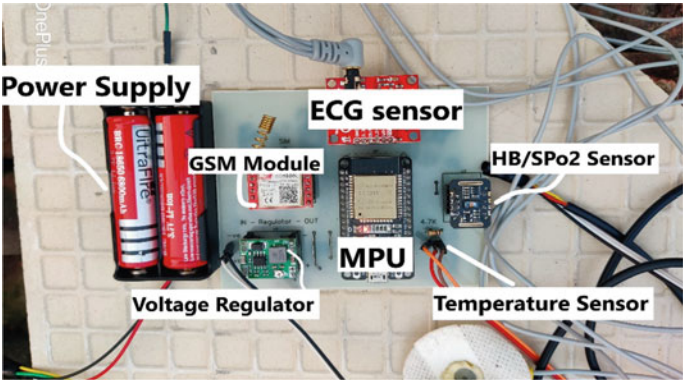


Fig. 6 Working “Smart Health Monitoring” device

After finalization of the circuit design and working code, a circuit diagram was designed using the EasyEDA PCB designing software. The designed PCB can be seen in Fig. 6.

The data collected from the connected devices are uploaded into Ubidots IoT platform which is available for the patient and the doctor to be monitored remotely.

Emergency Alert System. An emergency alert system has been imposed in the proposed system to notify the doctor and the patient in case of abnormal health data. Here, the abnormal health data refer to the data when it exceed the threshold value determined by the medical professional. The threshold value set in our system can be seen in Table 2.

During each cycle of the measurement of health parameters by the system, the parameters given in Table 2 are compared with the threshold values. Any value outside the maximum and minimum value are considered abnormal values and in case of two parameters showing abnormal value, an emergency text alert message is sent to the corresponding doctor and the patient. The algorithm is given in Fig. 7.

Table 2 Minimum and maximum threshold for HB, body temperature, and SpO2

Health parameters	Minimum value	Maximum value
Heart beat	60	100
Body temperature	36	38
Blood oxygen level SpO2	95	100

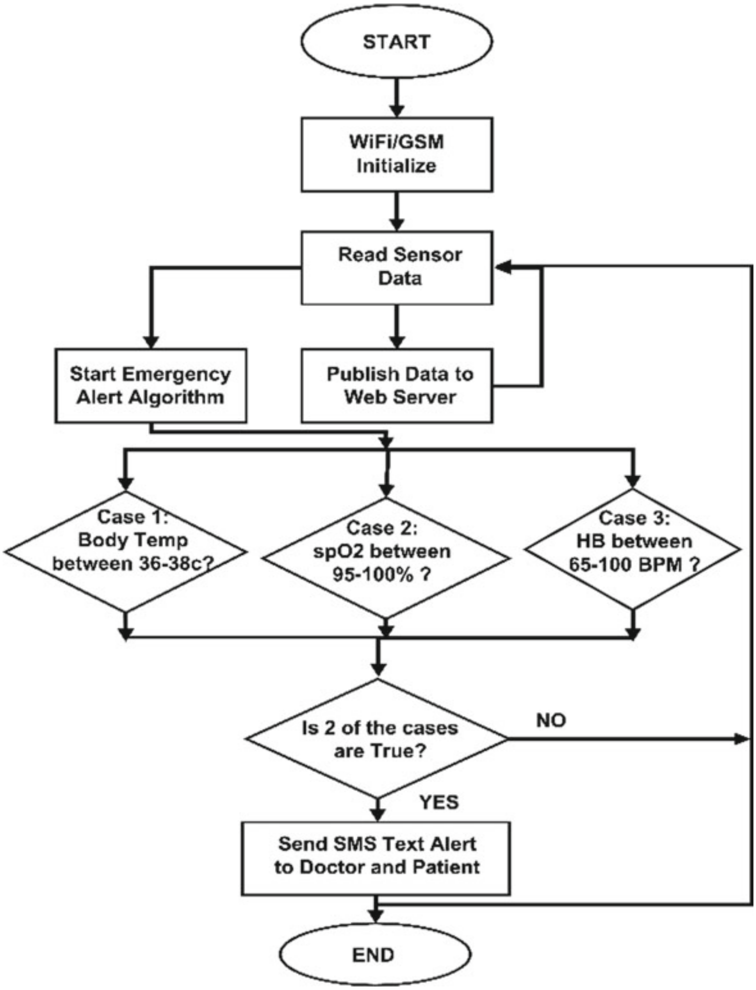


Fig. 7 Emergency alert system algorithm and flowchart

4 Results

The experiment was done on a subject where the sensors were connected to the subject, and the sensor data were monitored using the Ubidots IoT platform. In Fig. 8, we can see the completed working device.

Fig. 8 Working smart healthcare device

4.1 *Data Monitoring*

The ECG, body temperature, SpO2 levels, and heart beat per minute are published into MQTT server-based cloud service Ubidots. The published data are shown in Fig. 9.

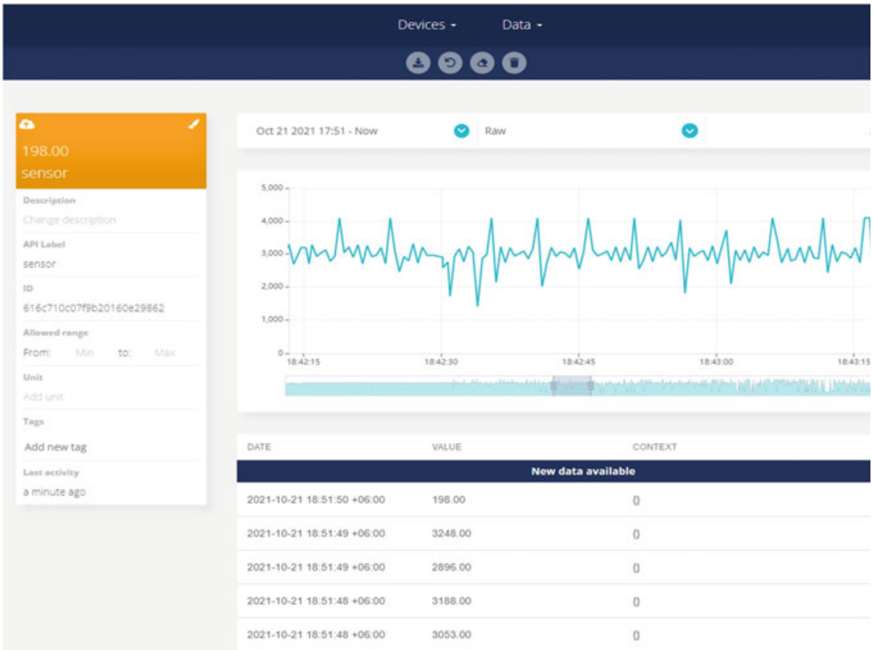


Fig. 9 Real-time data of vital health parameters and ECG shown in dashboard

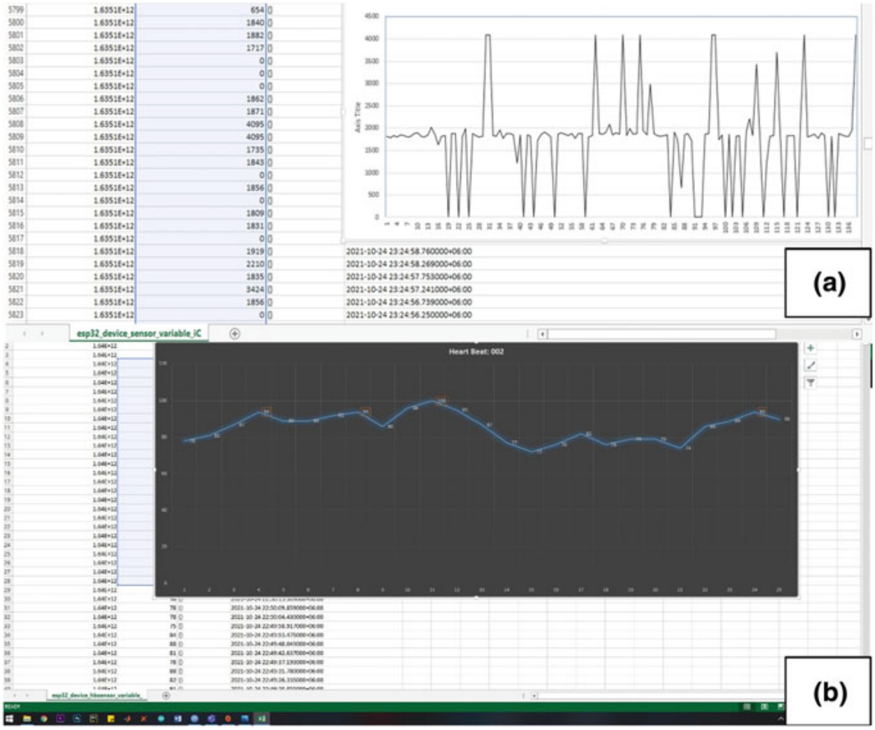


Fig. 10 ECG data (a) and heart rate data (b) of patient 002 in last 24 h

4.2 Data Analysis

The real-time health data can be monitored by medical professionals at any given time. But, in case of multiple patients, the health cannot be monitored by a single doctor at a time. Moreover, there are a lot of patients who are more critical than the other and need special attention. To overcome these problems, health data storage system has been used. The health data measured by the device are uploaded and stored in the Ubidots cloud server that can be accessed any time by the authority for further or in-depth inspection of the health parameter. We can see in Fig. 10 the ECG data and the heart beat of a patient (Patient ID: 002) are downloaded and inspected at a specific time window.

4.3 Emergency Alert System

The emergency alert system works according to the algorithm shown in Fig. 7. The threshold can be pre-determined by the doctor during the system set up. According to

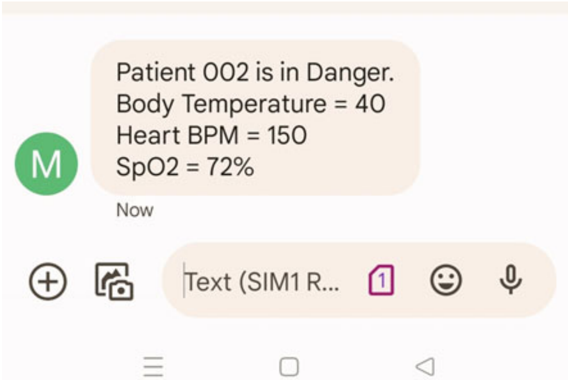


Fig. 11 Emergency alert for patient 002

the algorithm, two out of three parameters have to measure abnormal values in order to initiate an emergency alert. In this experiment, we have taken abnormal values of body temperature at 40 °C, heart rate of 150 bpm and SpO2 levels of 72%. The emergency alert message is seen in Fig. 11.

4.4 Data Accuracy

We have analyzed the average data of heart beat rate, blood oxygen levels, and body temperature from the stored data section of the IoT platform for various time windows and compared the data with store bought medical devices in order to calculate the data error percentage. We continued to collect and compare data of proposed system and medical devices of a 15 s, 30 s, 1 min, and 5 min time frame in order to properly justify the rate of error. The acceptable rate of heart beat is ±5% [20]. The average deviation of measured body temperature from actual temperature is 0.1–1.7 °C depending upon the point of measurement [21]. The error rate of pulse oximeter is about 3.6% as seen in [22]. The results of the experiments are given in Table 3.

As seen in Table 3, the rate of error of the measured values of vital health parameters by the proposed smart health monitoring system is under the acceptable rate defined by previous researches.

5 Discussion and Conclusion

The developed smart health monitoring system is a low-cost and efficient device that can sense and measure various types of health parameters such as heart beat rate, body temperature, blood oxygen levels known as SpO2 levels, and electrocardiodiagram known as ECG. The main purpose of this device is to upload the real-time health data to the Internet using the IoT platform which enables the doctor, the patient and anyone

Table 3 Comparison of actual data of medical instruments and monitored data of the system

Health parameters	Heart beat		Body temperature		Blood oxygen SpO2	
Acceptable rate	±5%		<4.5%		<3.6%	
	Actual	Monitored	Actual	Monitored	Actual	Monitored
First 15 s	78	72	37.4	37.0	99	97
Error%	4.87%		1.08% (37.0)		2.06% (97)	
	Actual	Monitored	Actual	Monitored	Actual	Monitored
First 30 s	76	74	37.5	37.3	99	97
Error%	2.70%		0.53%		2.06%	
	Actual	Monitored	Actual	Monitored	Actual	Monitored
First 1 min	80	83	37.5	37.5	98	96
Error%	3.75%		0%		2.08%	
	Actual	Monitored	Actual	Monitored	Actual	Monitored
First 5 min	84	80	37.5	37.4	99	98
Error%	5%		0.26%		1.02%	

with access, real-time remote monitoring capability. To measure the body temperature, we have used DS18B20, for heart rate and SpO2, we have used MAX30102 and for ECG, we have used AD8232 sensors. For IoT platform, Ubidots has been used. ECG monitoring can help the elderly people who needs continuous heart monitoring, blood oxygen level sensor helps with measurement of SpO2 which is a vital health parameters for COVID- and dengue-affected patient. Body temperature is the most common health parameter which is measured using the body temperature sensor. Our proposed and developed system minimizes work load of a doctor, health risk of the patient and saves valuable time of the users. Our future work includes emergency two-way communication system between the patient and the doctor. Not only that, but also in our future design, we will focus more on the portability of the device and also integrating more health sensors to the system. For example, glucometer for blood sugar measurement, sphygmomanometer for blood pressure measurement, etc. Also, an independent MQTT server-based web application and corresponding mobile application will be developed in the future.

References

1. Coronary heart disease in Bangladesh. World Life Expectancy. https://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease. Accessed 20 Oct 2021
2. Lung disease in Bangladesh. World Life Expectancy. https://www.worldlifeexpectancy.com/bangladesh-lung-disease. Accessed 20 Oct 2021
3. Prabhavathi, K., Selvi, K.T., Poornima, K.N., Sarvanan, A.: Role of biological sex in normal

cardiac function and in its disease outcome—a review. J. Clin. Diagn. Res. JCDR **8**(8), BE01–BE04 (2014). https://doi.org/10.7860/JCDR/2014/9635.4771
4. What is the normal body temperature: babies, kids, adults, and more. https://www.healthline.com/health/what-is-normal-body-temperature#temperature-range. Accessed 18 Apr 2021
5. Blood oxygen levels: what's all the hype about? Hackensack Meridian Health. 28 Sep 2020. https://www.hackensackmeridianhealth.org/HealthU/2020/09/28/blood-oxygen-levels-whats-all-the-hype-about/. Accessed 21 Oct 2021
6. Akash, M.R.R., Yousuf, Shikder, K.: IoT based real time health monitoring system. In: 2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH), Feb 2020, pp. 167–171. https://doi.org/10.1109/INBUSH46973.2020.9392163
7. Islam, M.M., Rahaman, A., Islam, M.R.: Development of smart healthcare monitoring system in IoT environment. SN Comput. Sci. **1**, 1–11 (2020)
8. Ruman, M.R., Barua, A., Rahman, W., Jahan, K.R., Roni, M.J. and Rahman, M.F.: IoT based emergency health monitoring system. In: 2020 International Conference on Industry 4.0 Technology (I4Tech), Feb 2020, pp. 159–162. https://doi.org/10.1109/I4Tech48345.2020.9102647
9. Reddy D.L., Naik M.R., Srikar, D.: Health monitoring system based on IoT. In: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), June 2021, pp. 468–472. IEEE
10. Bakar, N.H.B.A., Abdullah, K., Islam, M.R.: Wireless smart health monitoring system via mobile phone. In: 2016 International Conference on Computer and Communication Engineering (ICCCE), July 2016, pp. 213–218. IEEE
11. Yeri, V., Shubhangi, D.C.: IoT based real time health monitoring. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), July 2020, pp. 980–984. https://doi.org/10.1109/ICIRCA48905.2020.9183194
12. Valsalan, P., Tariq, A., Hussain, A.: IoT based health monitoring system. 2020. https://doi.org/10.31838/jcr.07.04.137
13. A smart health care monitor system in IoT based human activities of daily living: a review. https://ieeexplore.ieee.org/document/8862439. Accessed 21 Oct 2021
14. Rodrigues, E.M.G., Godina, R., Cabrita, C.M.P., Catalão, J.P.S.: Experimental low cost reflective type oximeter for wearable health systems. Biomed. Signal Process. Control. **31**, 419 (2017)
15. Hsia, B., Wang, D., Shaghaghi, N., Zhang, Z.: BOxy: a cost-effective blood oximeter
16. Ahmed, M.F., Hasan, M.K., Shahjalal, M., Alam, M.M., Jang, Y.M.: Design and Implementation of an OCC-based real-time heart rate and pulse-oxygen saturation monitoring system. IEEE Access **8**, 198740–198747 (2020)
17. Zhao, X., Li, W., Zhou, L., Song, G.B., Ba, Q., Ou, J.: Active thermometry based DS18B20 temperature sensor network for offshore pipeline scour monitoring using K-means clustering algorithm. Int. J. Distrib. Sens. Netw. **9**(6), 852090 (2013)
18. Olalekan, O.B.: Development of a Sim800l based reprogrammable household smart security system with recipient phone call alert. Int. J. Comput. Eng. Res. Trends **4**(1), 15–20 (2017)
19. Ubidots. https://iot-fpms.fandom.com/wiki/Ubidots. Last accessed 21 Oct 2021
20. Nelson, B.W., Allen, N.B.: Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. JMIR mHealth uHealth **7**(3), e10828 (2019). https://doi.org/10.2196/10828
21. Sund-Levander, M., Grodzinsky, E., Loyd, D., Wahren, L.K.: Errors in body temperature assessment related to individual variation, measuring technique and equipment. Int. J. Nurs. Pract. **10**(5), 216–223 (2004)
22. Report: Higher error rate seen in use of pulse oximeters with black patients. https://www.massdevice.com/report-higher-error-rate-seen-in-use-of-pulse-oximeters-with-black-patients/. Last accessed 22 Dec 2020

Smart Signals and NLP Agriculture

Toward Devising a Soil Parameters Monitoring System to Improve Plant Irrigation

Khadiza Newaz, Afsana Akhter, Umama Tasnim Tanisha, Md. Harunur Rashid Bhuiyan, Tarik Reza Toha, and Shaikh Md. Mominul Alam

Abstract The amount of arable land is shrinking in the world as the world's population is increasing. Producing more crops on less land now holds a significant importance in ensuring food security for this growing population. Soil parameters like soil pH, soil moisture, temperature, and humidity play a significant role in precision agriculture. Wrong land selection, over rainfall, and ignorance about maintaining appropriate soil parameters are the main obstacles in achieving a high yield of crops. A proper low-cost soil parameters monitoring system is yet to be proposed to the best of our knowledge. Therefore, automation of soil parameters monitoring can help in achieving an adequate yield of crops by making appropriate decisions. The objective of this paper is to propose a low-cost soil parameters monitoring system that monitors soil moisture, soil pH, along with environmental temperature wirelessly. Our proposed device collects soil parameters data from the agricultural field and then transfers these data to clients' device using a wireless network. To evaluate our device, we collect data from different plants and analyze the sensor data.

Keywords Soil parameters · Wireless · Soil moisture · Agriculture

1 Introduction

Crop production takes about 40% of the earth's land and consumes 85% of freshwater available in the world [20]. According to the United Nations, the global population is anticipated to grow from 7.3 billion in 2016 to 9.7 billion in 2050 [21]. Meeting global demand for safe and healthy food for the ever-increasing population now and into the future is currently a crucial challenge [37] for all countries. But the constant shrinking of arable land due to urbanization has made it even more challenging to meet this demand.

Improvement in crop yield can be achieved with increased profit margin and reduced pollution by performing better infield management decisions, reducing of

K. Newaz (✉) · A. Akhter · U. T. Tanisha · Md. H. R. Bhuiyan · T. R. Toha · S. Md. M. Alam
Bangladesh University of Textiles, Dhaka 1208, Bangladesh
e-mail: khadizanewaz2801@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_44

excess chemical and fertilizer costs, improved and efficient application of information technology, and permitting more accurate farm records [32]. In the field of precision agriculture, it is important to continuously monitor the fields as they are site specific. Precision agriculture requires permanent monitoring of the agricultural fields and intelligent processing of the measured data collected from the field, correlated with the weather forecasts [27]. Soil parameters like soil pH, soil moisture, temperature, and humidity play a significant role in obtaining high yields of crops [36].

In most countries, wrong land selection, insufficient rainfall, and ignorance about maintaining appropriate soil parameters are the main obstacles in precision farming. Despite its massive importance as human food provider and important role player in the international economy, the agriculture sector has been far behind than any other sector when it comes to applying advanced technology into it [19]. Backward agricultural systems and infield management are incapable of implementing advanced technologies in agriculture. As the world tends to new technologies and implementations, it is also necessary to implement them in agriculture [18]. Previous experience says that the application of advance technology in agribusiness facilitates the way for decision making, optimizes resources, and reduces the production cost. Nowadays, wireless sensor network (WSN) is worldwidely being used as a possible solution for precision agriculture. The wireless sensor network is being used to test the land to assess its suitability for cultivation and ensure that it is free from diseases and harmful fungi for old used lands, to do irrigation scheduling and fertilization scheduling [1] and to divert animal intrusions in the crop field for crop protection [7]. The monitoring of soil parameters with environmental indicators by wireless sensor network (WSN) will allow time and cost minimization and enable agriculture productivity to be maximized [16]. Therefore, we should focus on automation of soil parameters monitoring as well as environmental indicators monitoring, so that the status of agricultural land can be monitored at any time from anywhere and necessary decisions can be made accordingly.

In this paper, our objective is to propose a useful and efficient soil parameter monitoring system to help agricultural officials. In addition to the continuous monitoring of agricultural land, they can make appropriate decisions using the information collected by our system for time-saving and effective resource (water, fertilizer, pesticides) utilization that is conducive to achieving a high yield of the crop.

Based on our work, we make the following contributions to this paper:

1. We designed a low-cost, nRF-based, easy-to-implement soil parameters monitoring system that collects and transfers soil parameters data without using Wi-Fi or IoT for agricultural infield management.
2. We have been able to add more features to our device using relatively fewer sensors and hardware equipment than existing similar devices.
3. We deploy our device in both outdoors and indoors to collect real-time data of soil parameters.

The rest of this paper is organized as follows: In Sect. 2, the current state of the art is discussed, and Sect. 3 describes the methodology of the proposed wireless soil parameters monitoring system. In Sect. 4, the experimental evaluation is discussed.

The experimental evaluation includes deployment and testing of the proposed system in real-time data collection area, soil parameters analysis, implementation cost estimation, and summary of results, and Sect. 5 concludes the paper.

2 Related Work

Many pieces of research have already been done by other researchers on real-time soil parameters monitoring. Deep et al. [30] proposed an automated irrigation system. It monitors soil moisture, soil pH, and water flow from pump to land through a pipeline. It uses the concept of IoT to transmit the data to the website through the Internet. Payero et al. [24] researched wireless communication systems for monitoring soil moisture using Decagon EC-5 sensors. This system uses Arduino-compatible microcontrollers and communication systems to sample and transmit values from four Decagon EC-5 soil moisture sensors. Farooq et al. [10] proposed a low-cost smart crop monitoring and field irrigation system based on the IoT and mobile applications using soil moisture, temperature, humidity, gas and smoke, rain, PIR, and LDR sensors. This system uses Bluetooth module HC-05 or GSM module to transmit sensor values to farmers' cell phones through mobile applications. Tolentino et al. [38] proposed a system capable of acquiring temperature, moisture, pH level, and N-P-K values of the soil simultaneously to track soil fertility. Data was acquired wirelessly for versatile accessibility and with no delay through a Wi-Fi module connected to a wireless device. Chunduri et al. [8] introduced a smartphone and IoT-based real-time agricultural monitoring and controlling system that uses a temperature sensor, humidity sensor, light sensor, soil moisture sensor, soil pH sensor, and camera module for data acquisition. Prasad et al. [28] proposed an agricultural autonomous system that will sense the field conditions in real time and analyze the field parameters such as temperature, soil moisture, and humidity. The collected data is transferred to the LoRa Gateway and then sent to the cloud server via Wi-Fi or Ethernet. Islam et al. [14] proposed an IoT-based real-time agro field monitoring and irrigation controlling system using ambient temperature and humidity sensor, soil moisture sensor, and soil temperature sensor. In this system data from the sensor, the node is transmitted to the cloud server using LoRa IoT protocol through the gateway node. Rao et al. [31] introduced an IoT-based automatic agricultural field control and real-time data monitoring system. This system mainly focuses on moisture variations correlated with temperature changes data by smart sensors and controls irrigation system. Islam et al. [15] introduced a smart device for an irrigation system that is also capable of monitoring the temperature, moisture, and humidity of the crop field and the saltwater intrusion by using temperature, moisture, humidity sensor, and salinity sensor. Ramson et al. [29] proposed a system for the development, deployment, and validation of an IoT system to monitor the dynamic soil properties in real time and to ease data collection in remote field sites. Dorji et al. [9] designed an electronic nose (e-nose)-based wireless sensor network to monitor soil nutrients.

In this system, Zigbee connected to Arduino Nano is used to transmit collected data wirelessly. This system operates easily in real time with a battery charged using a solar cell. Saqib et al. [33] introduced a soil moisture information monitoring approach for collecting field data over long distances that can be used in a fully automated agricultural farm. Ananthi et al. [6] bring in an IoT-based system for soil monitoring and irrigation to reduce the manual monitoring of the field and get the information via mobile application. In this methodology, obtained sensor values are sent to the field manager through the Wi-Fi router, and the crop suggestion is made through the mobile application. Patil et al. [23] proposed an integrated system for farm monitoring based on smartphone and IoT technology using soil moisture sensor, DHT temperature sensor, leaf wetness duration (LWD) sensor, and pH sensor. This system continuously collects real-time data from sensors and transmits these data to farmers' device using both android and web platforms.

In our literature survey, we found that most of the research papers on real-time soil parameters monitoring need an uninterrupted Internet connection to transfer their collected sensor data to the client device and the cloud server. Many of the works used solar panel-based power source charging systems or used high power-consuming expensive modules. Some proposed systems were able to transfer data up to a distance of only a few feet. On the other hand, some of the proposed devices were very expensive which could discourage the use of this device at initial stage. In this paper, our objective is to implement a low-cost, nRF-based easy hardware system for continuous soil parameters monitoring that will facilitate the way of precision agriculture. As there are no uninterrupted Internet facilities in rural agricultural areas and solar panel-based power sourcing systems are uncertain in the rainy season, using expensive high power-consuming networking modules has made the above-mentioned systems much complex and costly.

Therefore, these approaches are not readily applicable in rural agricultural fields. Hence, we need to devise a customized soil parameters monitoring system using low-cost hardware resources (Fig. 1).

3 Proposed Methodology

Our proposed device consists of two units: the sender unit and the receiver unit. The sender unit collects soil parameters data by sensing modules, and the controller module sends the data via nRF to the receiver unit. The receiver unit receives the information via another nRF and stores it on the laptop using a software. Besides, the receiver unit displays the data in a smartphone by sending the data via Bluetooth. The block diagram of our proposed device is shown in Fig. 2. Next, we describe the units of our device in detail.

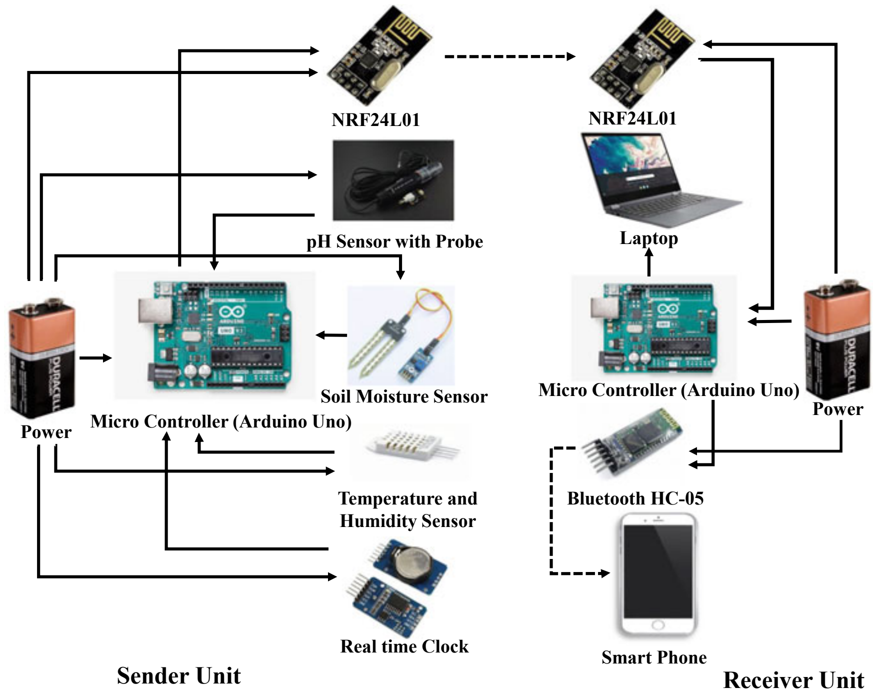


Fig. 1 Block diagram of our proposed system

3.1 Sender Unit

The sender unit consists of various types of sensors, such as soil moisture sensor, pH sensor, temperature, and humidity sensor. The microcontroller controls these modules. We used the Arduino Uno R3 for developing our device, which can control our entire system. It operates every sensor from 3.3 to 5 V. Soil moisture sensor SEN-00245 is used to measure the volumetric water content of the soil. Gravity Analog pH sensor-V2-SEN0161 is used to measure the pH of the soil. The DHT22 digital temperature and humidity sensor consumes low power and provides a digital output. We also used a DS-3231 RTC module to collect data in real time. DS-3231 is an accurate 12C real-time clock (RTC). It has an integrated temperature-compensated crystal oscillator (TCXO) and crystal.

3.2 Receiver Unit

The receiver unit consists of an Arduino Uno R3, nRF24L01 module. The sensor data is received by nRF24L01 module. Then the data is saved on the laptop using

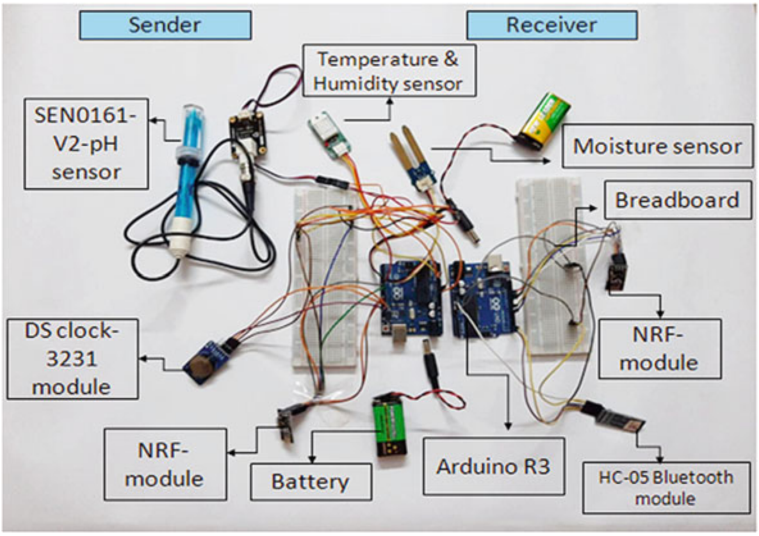


Fig. 2 Experimental setup of our proposed methodology

PuTTY software on a continuous process. By adding a Bluetooth HC-05 module, the data can be also displayed on a smartphone.

4 Experimental Evaluation

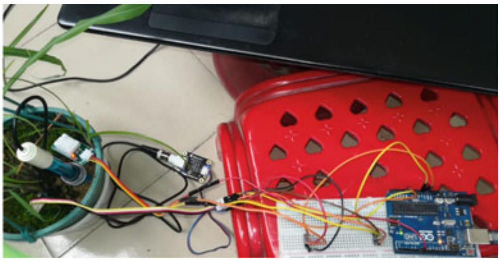
To assess our proposed system in the real world and to collect real-time soil parameters data, we deploy our proposed soil parameters monitoring system in both open environments and indoors. Figure 2 shows our experimental setup used in a real-world deployment. Next, we describe the data collection process of different sensors used in our system in detail.

4.1 Outdoor Data Collection

To collect open environmental data with our device, we briefly observed the soil parameters of several plants growing in open space for several days. Our outdoor data collection process is shown in Fig. 3a.

Soil Moisture Sensor We instantaneously monitored the change in soil moisture every second using the serial monitor. The result of our soil moisture monitoring in an open environment is shown in Fig. 4a. It is noticed from the resulting values of sensor data that soil moisture percentage increases with the amount of water in the soil. We noticed that the sensor gives lower values in comparatively dry soils. We also

(a) Outdoor data collection



(b) Indoor data collection

Fig. 3 Snapshots of real-time data collection

noticed that there is a huge difference between measured soil moisture percentage of different plants and ideal values [2, 22, 25, 34] of different plants due to improper irrigation planning.

pH Sensor We monitored the change in soil pH of different plants with our pH sensor. Our observations shown in Fig. 4c present that soil pH sensor data varies from plant to plant considerably, which sometimes does not match the ideal range [5, 11, 13, 17]. The main reasons for this violation of soil pH from the ideal range are wrong land selection, improper fertilization plan, and soil pollution. Our sensor data also shows that some herbs (Coriander) can thrive in slightly acidic soil and some climbing plants (Indian Spinach) can grow well even in mild alkaline soil.

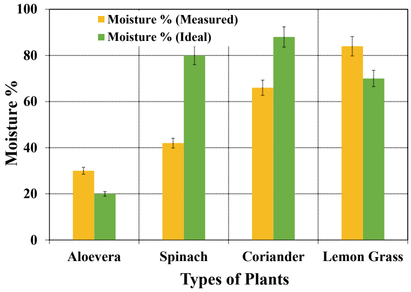
Temperature and Humidity Sensor We observed the rise and fall of soil temperature continuously with our DHT22 sensor. Our observation in Fig. 4d shows that the environmental temperature value provided by the sensor is higher when the moisture sensor gives the lower value. It is also seen from the graph in question that optimum environmental temperature differs from plant to plant, which sometimes does not match the ideal value [3, 4, 12, 26].

4.2 Indoor Data Collection

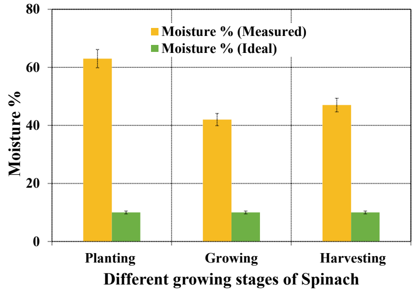
We collected data from the Indian Spinach plant as an indoor plant. Figure 3b shows our indoor data collection process. Our observation in Fig. 4b shows the variation of soil moisture percentage of Indian Spinach in different growing stages in indoor environment. Data collected from Indian Spinach shows that planting and harvesting stages require more water in the soil than growing stage.

4.3 Cost Estimation

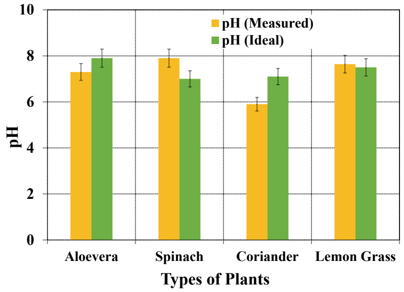
Throughout the designing process of our proposed system, we tried our best to keep the total cost of our device as low as possible so that our device can easily reach even the farmers of rural areas.



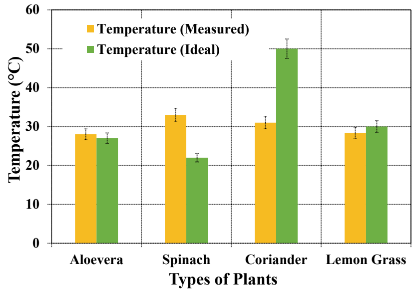
(a) Soil moisture% of different plants



(b) Soil moisture% of Indian Spinach in different growing stages



(c) pH of soil for different plants



(d) Environmental temperature for different plants

Fig. 4 Different sensor data collected in open and indoor environment

In this part of the paper, we are comparing the implementation cost of our device with the implementation cost of some other existing similar devices [24, 29]. Of the two existing devices mentioned in the reference, the implementation cost of the LoRaWAN IoT-based [29] device proposed by Ramson et al. is 2477.52 USD and the implementation cost of the Arduino-based wireless [24] device proposed by Payero et al. is 610 USD. The implementation cost analysis of our proposed system is given in Table 1. The price details of each piece of hardware equipment were collected from Techshopbd [35].

4.4 Findings

This paper has clearly shown that soil parameters change very often. So, constant and careful monitoring of cultivable land is needed to get the maximum yield of crops. According to the graphs plotted with the values of sensor data shown in Fig. 4, it is evident that soil parameters vary by plants, planting environments, temperature, and even specific stages of harvesting.

Table 1 Cost analysis of our proposed system

Component name	Model name	Quantity	Unit price (USD)
Microcontroller	Arduino Uno R3	2	9.47
Breadboard		2	1.33
Real-time clock	DS–3231	1	2.66
Temperature and humidity sensor	DHT22	1	6.125
Soil moisture sensor	SEN-00245	1	2.06
pH sensor	V2-SEN0161	1	52.325
NRF24L01	PCB Antenna	2	2.5
Bluetooth module	HC-05	1	12.41
Power supply	9V battery	2	0.70
Total cost			105 USD

It is observed from Fig. 4a that soil moisture percentage variation from the ideal value is excessively higher than any other parameters. The reasons behind this variation of optimum soil moisture are unplanned irrigation systems, inadequate rainfall, and unpredictable temperature of the summer-monsoon season. It is also seen from Fig. 4b in the case of Indian Spinach that soil moisture percentage demand of plants changes throughout different stages of growing (planting to harvesting). pH sensor data shown in Fig. 4c says that optimum soil pH is not maintained due to improper fertilization plan, wrong land selection, and soil pollution. Figure 4d shows that the measured environmental temperature lies within the optimum range in few cases but is also found to violate the ideal value in other cases. Following our observations, the main reason for this irregular fluctuation in environmental temperature is global warming and climate change.

Hence, soil parameters data can be used to monitor agricultural land overcoming the limitations of round-the-clock monitoring of agricultural land due to distance and adverse weather conditions and lead to high and quality yield of crops.

5 Conclusion

Wireless soil parameters monitoring device can play an important role in the development of the plant irrigation system. Existing other devices of the same purpose have a more complex and costly architecture which made their initial implementation limited. Addressing these limitations, this paper proposes a low-cost nRF-based wireless soil parameters monitoring system to improve plant irrigation. By installing this device in the agricultural field, the agriculture officials while sitting in their office as well as the farmer himself can observe the current condition and immediate changes of soil parameters in the agricultural land from anywhere in the world.

Real-time soil parameters data monitoring with our proposed device will facilitate agricultural officials to take the necessary steps to make appropriate decisions to achieve the maximum yield of crops. Implementation of this device and deployment in agricultural field will also overcome the limitations of round-the-clock monitoring of agricultural land due to distance and adverse weather conditions. In this experiment, sample plants for data collection are selected in consideration of the COVID-19 situation as we were not able to go to the agricultural field for data collection.

In the future, we will work to collect real-time agricultural field data of more wide range with multiple sensor nodes. We will also work to use our device to improve cotton and jute field irrigation system.

Acknowledgements We give our gratitude to Bangladesh University of Textiles for providing financial support to this research work.

References

1. Abd El-kader, S.M., El-Basioni, B.M.M.: Precision farming solution in Egypt using the wireless sensor network technology. Egypt. Inf. J. **14**(3), 221–233 (2013)
2. Alkimim, E.R., David, A., Sousa, T.V., Rodrigues, C.G., Amaro, H.T., et al.: Different harvest times and physiological quality of coriander seeds. Rev. Bras. de Engenharia Agrícola e Ambiental **20**, 133–137 (2016)
3. Allthatgrows: How to Grow Spinach at Home in India. https://www.allthatgrows.in/blogs/posts/grow-spinach-at-home-in-india (2021). Accessed on 15 Oct 2021
4. Almanac: Aloe Vera: How to Care for Aloe Vera Plants | The Old Farmer's Almanac. https://www.almanac.com/plant/aloe-vera (2021). Accessed on 15 Oct 2021
5. Almanac: Optimum Soil ph Levels for Plants | The Old Farmer's Almanac. https://www.almanac.com/plant-ph (2021). Accessed on 15 Oct 2021
6. Ananthi, N., Divya, J., Divya, M., Janani, V.: Iot based smart soil monitoring system for agricultural production. In: 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 209–214. IEEE (2017)
7. Bapat, V., Kale, P., Shinde, V., Deshpande, N., Shaligram, A.: WSN application for crop protection to divert animal intrusions in the agricultural land. Comput. Electron. Agric. **133**, 88–96 (2017)
8. Chunduri, K., Menaka, R.: Agricultural monitoring and controlling system using wireless sensor network. In: Soft Computing and Signal Processing, pp. 47–56. Springer (2019)
9. Dorji, U., Pobkrut, T., Kerdcharoen, T.: Electronic nose based wireless sensor network for soil monitoring in precision farming system. In: 2017 9th International Conference on Knowledge and Smart Technology (KST), pp. 182–186. IEEE (2017)
10. Farooq, M.U., Hakim, A., Baig, I.A., Khanna, P., Jabeen, J., Ahmed, U.I.: Low-cost smart crop monitoring and irrigation system based on iot and mobile application. Open J. Soil Sci. (2021)
11. Gardenexpertguide: How to Grow Coriander in Australia? - Garden Expert Guide. https://www.gardenexpertguide.com/grow-coriander-in-australia/ (2021). Accessed on 15 Oct 2021
12. Garden expertguide: How to Grow Coriander in Australia? - Garden Expert Guide. https://www.gardenexpertguide.com/grow-coriander-in-australia/ (2021). Accessed on 15 Oct 2021
13. Homeguides.sfgate: Home Guides | sf gate. https://homeguides.sfgate.com/correct-ph-balance- (2021). Accessed on 15 Oct 2021
14. Islam, A., Akter, K., Nipu, N.J., Das, A., Rahman, M.M., Rahman, M.: Iot based power efficient agro field monitoring and irrigation control system: an empirical implementation in precision

agriculture. In: 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), pp. 372–377. IEEE (2018)
15. Islam, M.S., Dey, G.K.: Precision agriculture: renewable energy based smart crop field monitoring and management system using wsn via iot. In: 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1–6. IEEE (2019)
16. Jawad, H.M., Nordin, R., Gharghan, S.K., Jawad, A.M., Ismail, M.: Energy-efficient wireless sensor networks for precision agriculture: a review. Sensors **17**, 1781 (2017)
17. Johnnyseeds: Lemon Grass Plants (West Indian) | Key Growing Information. https://www.johnnyseeds.com (2021). Accessed on 30 Oct 2021
18. Krishnan, S., Kumaran, S.S., Preethi, M.R.: Field monitoring using wireless sensor networks based on iot. IRJET (2020)
19. Mentsiev, A.U., Amirova, E., Afanasev, N.: Digitalization and mechanization in agriculture industry. In: IOP Conference Series: Earth and Environmental Science. vol. 548, p. 032031. IOP Publishing (2020)
20. Nalina, K., Kaliwal, R.: Analysis of soil parameters in agriculture field using iot. Int. Res. J. Eng. Technol. 2977–2980 (2017)
21. Nicol, L., Nicol, C.: Adoption of precision agriculture to reduce inputs, enhance sustainability and increase food production: a study of southern alberta. Canada. WIT Trans. Ecol. Environ **217**, 327–336 (2018)
22. Oladele, O., Aborisade, A., et al.: Influence of different drying methods and storage on the quality of Indian Spinach (*basella rubra l.*). Am. J. Food Technol. **4**(2), 66–70 (2009)
23. Patil, M.A., Adamuthe, A.C., Umbarkar, A.: Smartphone and iot based system for integrated farm monitoring. In: Techno-Societal 2018, pp. 471–478. Springer (2020)
24. Payero, J.O., Nafchi, A.M., Davis, R., Khalilian, A., et al.: An arduino-based wireless sensor network for soil moisture monitoring using decagon ec-5 sensors. Open J. Soil Sci. **7**(10), 288 (2017)
25. Pisalkar, P., Jain, N., Jain, S.: Osmo-air drying of aloe vera gel cubes. J. Food Sci. Technol. **48**(2), 183–189 (2011)
26. Plantfoodathome: How to Grow Lemongrass: Planting and Growing Guide. https://www.plantfoodathome.com/grow-lemongrass/ (2021). Accessed on 30 Oct 2021
27. 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)
28. Prasad, S., Thangatamilan, M., Suresh, M., Panchal, H., Rajan, C.A., Sagana, C., Gunapriya, B., Sharma, A., Panchal, T., Sadasivuni, K.K.: An efficient lora based smart agriculture management and monitoring system using wireless sensor networks. Int. J. Ambient Energy, 1–10 (2021)
29. Ramson, S.J., León-Salas, W.D., Brecheisen, Z., Foster, E.J., Johnston, C.T., Schulze, D.G., Filley, T., Rahimi, R., Soto, M.J.C.V., Bolivar, J.A.L., et al.: A self-powered, real-time, Lorawan iot-based soil health monitoring system. IEEE Internet Things J. **8**(11), 9278–9293 (2021)
30. Rani, D., Kumar, N., Bhushan, B.: Implementation of an automated irrigation system for agriculture monitoring using iot communication. In: 2019 5th International conference on signal processing, computing and control (ISPCC), pp. 138–143. IEEE (2019)
31. Rao, R.N., Sridhar, B.: Iot based smart crop-field monitoring and automation irrigation system. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 478–483. IEEE (2018)
32. Sahu, B., Chatterjee, S., Mukherjee, S., Sharma, C.: Tools of precision agriculture: a review. Int. J. Chem. Stud **7**, 2692–2696 (2019)
33. Saqib, M., Almohamad, T.A., Mehmood, R.M.: A low-cost information monitoring system for smart farming applications. Sensors **20**(8), 2367 (2020)
34. Semanticscholar: Proximate and Phytonutrient Content of Cymbopogon Citratus (lemongrass) Leaf Extract and Preparation of Herbal Cookies | Semantic Scholar. https://www.semanticscholar.org/paper/Proximate-and-phytonutrient-content-of-Cymbopogon-Pp-Ar/216309d88d41ae95e2ae46bd8158e89b7fba218a (2021). Accessed on 30 Oct 2021

35. Techshopbd: Techshopbd.com | Largest Robotics-Electronics Shop in Bangladesh. https://www.techshopbd.com/ (2021). Accessed on 18 Oct 2021
36. Thakur, D., Kumar, Y., Kumar, A., Singh, P.K.: Applicability of wireless sensor networks in precision agriculture: a review. Wireless Personal Commun. **107**(1), 471–512 (2019)
37. Timsina, J.: Can organic sources of nutrients increase crop yields to meet global food demand? Agronomy **8**(10), 214 (2018)
38. Tolentino, E.V.N., Andaya, V.S., Cristobal, G.A.G., Ongtengco, R.S., Rosal, A.A., Ruzol, E.B., Sacramento, J.C.A.: Development of wireless data acquisition system for soil monitoring. In: IOP Conference Series: Earth and Environmental Science, vol. 463, p. 012088. IOP Publishing (2020)

DIIGMCS—Design and Implementation of IoT-Based Greenhouse Monitoring and Controlling System

Avizit Chowdhury Bappa, Moon Chowdhury, Golap Kanti Dey, and M. R. Karim

Abstract This paper reports a system called "DIIGMCS—design and implementation of IoT-based greenhouse monitoring and controlling system" designed and implemented for monitoring and controlling the environmental parameters to provide a required environment inside a greenhouse for small plants. The heater bulb, cooling fan, and pump motor are used to adjust the environment inside the greenhouse. IoT is implemented to monitor and control temperature, humidity, and soil moisture remotely by the Website. The project consists of two parts, namely hardware and software. For the hardware part, a system is modeled with several sensors that can measure the value of the moisture from the moisture sensor. The DHT11 sensor contains two sensors; one is temperature, and the other is humidity sensor. The DHT11 sensor gives us both temperature values and humidity values. These values which we get from the sensors, all are values. The proposed system is controlled by an Arduino, which is in turn interfaced with an LCD display as well as a Wi-Fi connection in order to transmit data. The system shows data of sensors over IoT in real time. This system has two different modes such as automatic and server mode to control the environmental parameters inside the greenhouse automatically by sensors threshold value or manually by using an IoT-based Website. Thus, the IoT-based greenhouse monitoring system effectively uses the internet to monitor sensor status and save plants live on time.

Keywords DIIGMCS · IoT · Greenhouse · Temperature · Humidity · Soil moisture

1 Introduction

Greenhouses are essential for growing plants such as flowers and vegetables. It is useful to protect crops from many diseases and insects. Many farmers fail to get good profit from the greenhouse because they cannot manage the essential factors

A. Chowdhury Bappa · M. Chowdhury · G. Kanti Dey · M. R. Karim (✉)
Department of Electrical and Electronic Engineering, Chittagong Independent University, Chattogram, Bangladesh
e-mail: mrkarim@ciu.edu.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_45

which control the plants growth and productivity. Greenhouse temperature should not go below a particular degree; high humidity may result in crop transpiration, condensation of water vapor on various greenhouse surfaces, and water evaporation from the humid soil. This can be possible using Internet of Things (IoT)-based wireless [1–3] sensors. To beat such challenges, IoT-based greenhouse monitoring [4] and controlling system involves the rescue. Through this system, we can remotely monitor and control the environmental parameters inside the greenhouse such as temperature, soil moisture, and humidity.

This greenhouse system is powered by an Arduino [5, 6] controlled microcontroller that consists of a temperature sensor, soil moisture sensor, LCD display module, 12 V DC cooling fan, 220 V 60 W heater bulb, and 12 V DC pump motor. The temperature sensor senses the extent of temperature, if it goes high DC fans get on and when the temperature goes low the fan gets off. Within the absence of sunshine, when the temperature goes low, the bulb starts glowing. Soil moisture sensor senses the water level because the level decreases as the pumps get on. During this way, it will become easy to real-time control and monitor the system.

Nowadays, the Internet has about-turned one of the compulsory corridors of our actuality. But the actual definition of IoT is making a brilliant, invisible network which may be perceived, controlled, and programmed. The network of physical objects “things” that are integrated with sensors, software, and other technologies described by the Internet of Things with the goal of connecting and exchanging data with other devices and systems via the web.

The concept of Internet of Things stands on real time, continued, remote observing, wearable sensors [7], long battery life of a designed device, and wireless network which enable users to interact and reach the application information.

The main purpose of the project is to remotely monitor the greenhouse assets in an innovative way to provide better, instantaneous, useful, and low-cost access to proper plant growing facilities. It gathers total information in real time of the essential factors which control the plants growth and productivity. In this project, data like temperature, soil moisture, humidity, etc., of the greenhouse is monitored by other users at remote locations through IoT [8–10].

The objective of the project is to implement the greenhouse [11–13] system by monitoring the temperature [14], humidity, and soil moisture using wireless communication. The system is based on an Arduino equipped with a set of sensors and wireless communication unit. Used sensors data is sent to Website through Internet connectivity. With the dual mode automatic and server mode, one can easily control the system automatically by sensors threshold value or manually from the IoT Website by switch on and off the loads such as heater bulb, cooling fan, or pump motor.

The aim of the project is limited to soil moisture [15–17], temperature, humidity [18], and remote viewing of the managed data from the sensors. The entire systems are controlled by Wi-Fi connected Arduino Nano and require some programming work. This project helps to monitor all data from sensors in real time using an IoT-based web server and also can control the parameters inside the greenhouse manually from Website by any user from anywhere. Greenhouse monitoring is labor-intensive and time-consuming without using Internet of Things. This proposed model saves time, money, and human effort.

2 Framework of the System

The proposed system architecture of Fig. 1 consists of two parts that is hardware and software. On the hardware section, the model is interfaced with various sensors and devices. Sensors will monitor environmental parameters such as temperature, humidity, and soil moisture inside the greenhouse which then transfer over the Internet to be entered by other users from a remote location. The system is organized by Wi-Fi module ESP8266 connected Arduino Nano and needs some programming codes. Through Wi-Fi module, it connected to the IoT web server. Arduino IDE software has been used for the software section. The system has two different modes as automatic and server mode. If the system mode is set to automatic, then the system will check the sensors value and calculate them. If the calculated value crosses the threshold value, then the loads pump motor, heater bulb, and cooling fan will be turned on. On the other hand, if the calculated value is not crosses the threshold value, then the system loads remain turned off. As the system consists both the hardware and software, the features of the hardware part need to be considered for the project implementation besides with the appropriate circuit connections and figures also the programming codes for the project functioning.

3 Constituent Sub-modules

3.1 Arduino Nano

The Arduino Nano as shown in Fig. 2 is supported by the ATmega328p which consists of 8 analog pins, 14 digital I/O pins, 6 power pins, and 2 reset pins. The operating voltage of this device is 5 V; however, the input voltage can vary between 7 and 12 V with maximum of 40 mA current and the crystal oscillator frequency of 16 MHz.

3.2 Wi-Fi Module (ESP8266)

The Wi-Fi module ESP8266 is an economical device to provide an Internet connection to systems. It works both as a hotspot or can connect with Wi-Fi network, so it can easily obtain information and transfer it to the web as easy as possible. It is often programmed using the Arduino IDE which makes it more convenient. Three

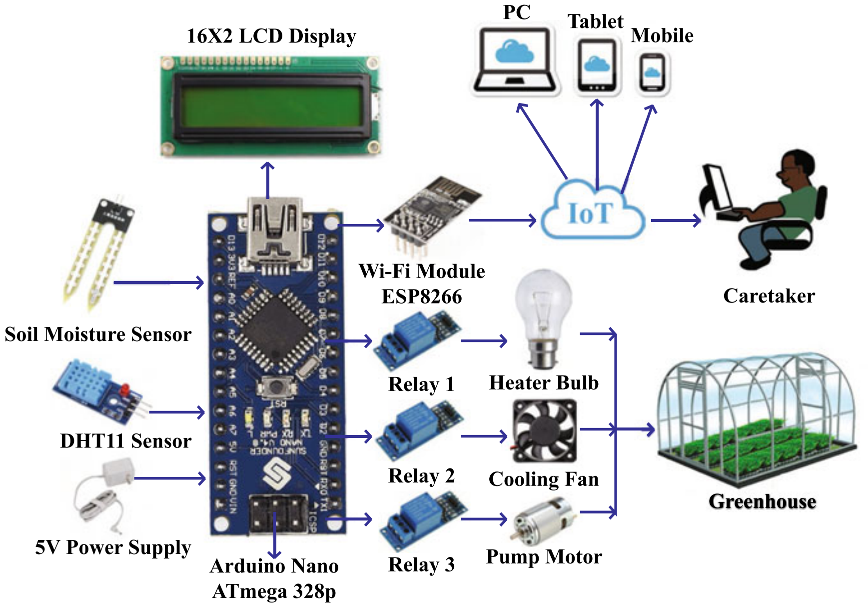


Fig. 1 Proposed architecture of DIIGMCS

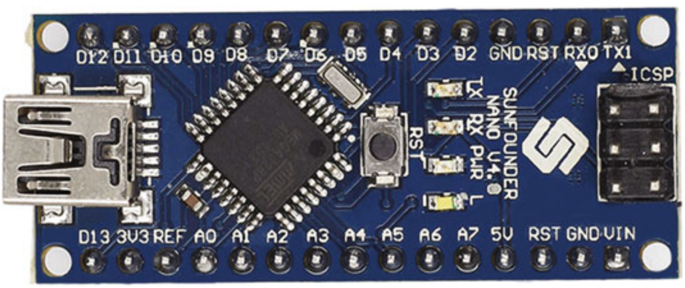


Fig. 2 Arduino Nano [19]

different criteria have been selected to observe the greenhouse environmental such as temperature, soil moisture, and humidity.

3.3 Sensor: Temperature and Humidity (DHT11)

DHT11 as shown in Fig. 3 is employed to detect temperature and humidity of the greenhouse environment. The humidity value from 20 to 90% and the temperature

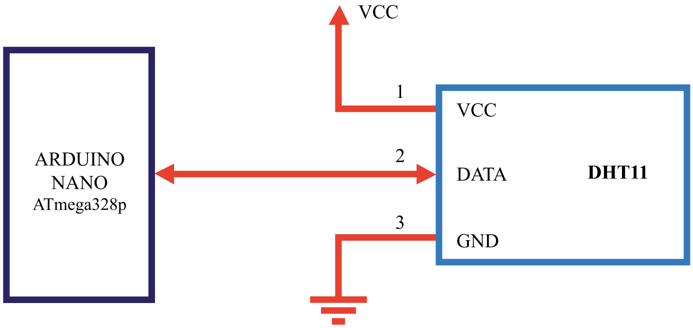


Fig. 3 Humidity and temperature sensor (DHT11) application circuit diagram

from 0 to 50 °C can be measured by this device with an accuracy of ± 1% and ± 1 °C.

3.4 Soil Moisture Sensor

The sensor shown in Fig. 4 is one of the crucial elements used to detect soil moisture in our proposed DIIGMCS system. Figure 5 shows the soil moisture sensor circuit which is mainly used to detect the moisture of the soil in case it is dry or not. It has two probes which work as electrodes to estimate the volumetric water content. When the soil is dry, it gives high resistance value which means less conductivity; otherwise, it shows low resistance value means high conductivity.

Fig. 4 Soil moisture sensor [20]

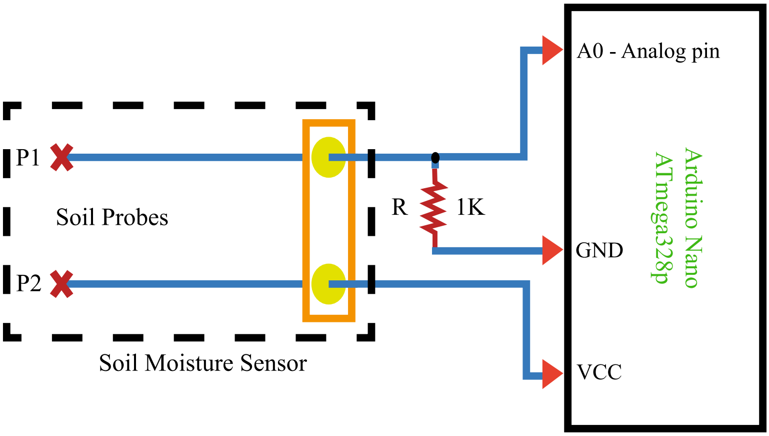


Fig. 5 Soil moisture sensor application circuit diagram

3.5 Load Elements

In this proposed DIIGMCS system, three load elements such as 220 V 60 W heater bulb, 12 V DC pump motor, and 12 V DC cooling fan are used to control the environmental factors such as temperature, humidity, and soil moisture. All three loads of the system controlled by switching on/off the 5 V DC relay module. The 5 V DC relay is programmed and controlled by Arduino Nano. Whenever the sensors cross the threshold values the microcontroller process the data and switch on/off the relays automatically according to the instruction. The load elements also can control manually from the IoT Website by simply switch on/off the buttons from the controlling panel. The current status of the relay switches also shown in the controlling panel of the IoT Website also on the LCD display.

3.6 Software Used

Arduino IDE is used to code and upload programs to Arduino Nano, and Sublime Text web code editor is used to develop the frontend and backend of the IoT Web site.

4 Methodology

In this project, DHT11 sensor is applied to detect and monitor the temperature and humidity of the greenhouse and the monitored data is sent to the Website using IoT.

This data shows on the IoT Web site as well as on the LCD display. By following the temperature and humidity threshold value of the sensor relay module, turn on/off the heater bulb and cooling fan. The values of the sensors data can real-time monitor or can control it manually by user from the IoT Web site. A soil moisture sensor is used to trace the moisture of the soil weather it is dry or not which will monitor the soil moisture data and sent them to the Web site using IoT.

This data shows on the IoT Web site as well as on the LCD display. By following the soil moisture threshold value of the sensor relay module, turn on/off the pump motor to supply water to the plants. Whenever the threshold value goes low, it means the soil is dry and the pump motor will automatically turn on by the system. When the threshold value goes high, it means the soil is wet and the pump motor will automatically turn off by the system.

Figure 6 represents the flowchart of overall project task of the IoT-based greenhouse monitoring and controlling system which executes into several steps. When we start the system, it shows two different modes such as automatic and server mode. So, we can control the system manually by server mode where we can turn on the pump motor, heater bulb, and cooling fan by switch on the relays remotely from IoT Web site. In automatic mode, the system will work automatically by following the sensors threshold value. All the environmental factors inside the greenhouse including temperature, humidity, and soil moisture are measured by the sensors and can monitor them through IoT server from remote place. Whenever the sensor crosses the threshold value, it will turn on/off the pump motor, cooling fan, and heater bulb to fulfill the environmental condition inside the greenhouse. The proposed system each step is monitored and can be controlled by using IoT Web site.

5 Result and Implementation

The proposed DIIGMCS system is designed in such manner that new users with limited knowledge of the web should be capable to access this application. The purpose of the project is continuous and effective distant monitoring and controlling of the parameters inside greenhouse in such a way to deliver preferred, instant, profitable, and economical access to suitable cultivation facilities.

Table 1 shows the data value of temperature, humidity, and soil moisture with respect to time and date. Figure 7 depicts the graphical representation of the temperature (degree C), humidity (%) and soil moisture (%) from the value (0 to 100) inside the greenhouse with respect to time.

A screenshot of the monitoring and controlling panel with output data of different sensors including humidity, soil moisture, and temperature on the IoT Web site is shown in Fig. 8.

The values of humidity, temperature, and soil moisture of the proposed greenhouse can be monitored by other users by staying at a remote location through IoT.

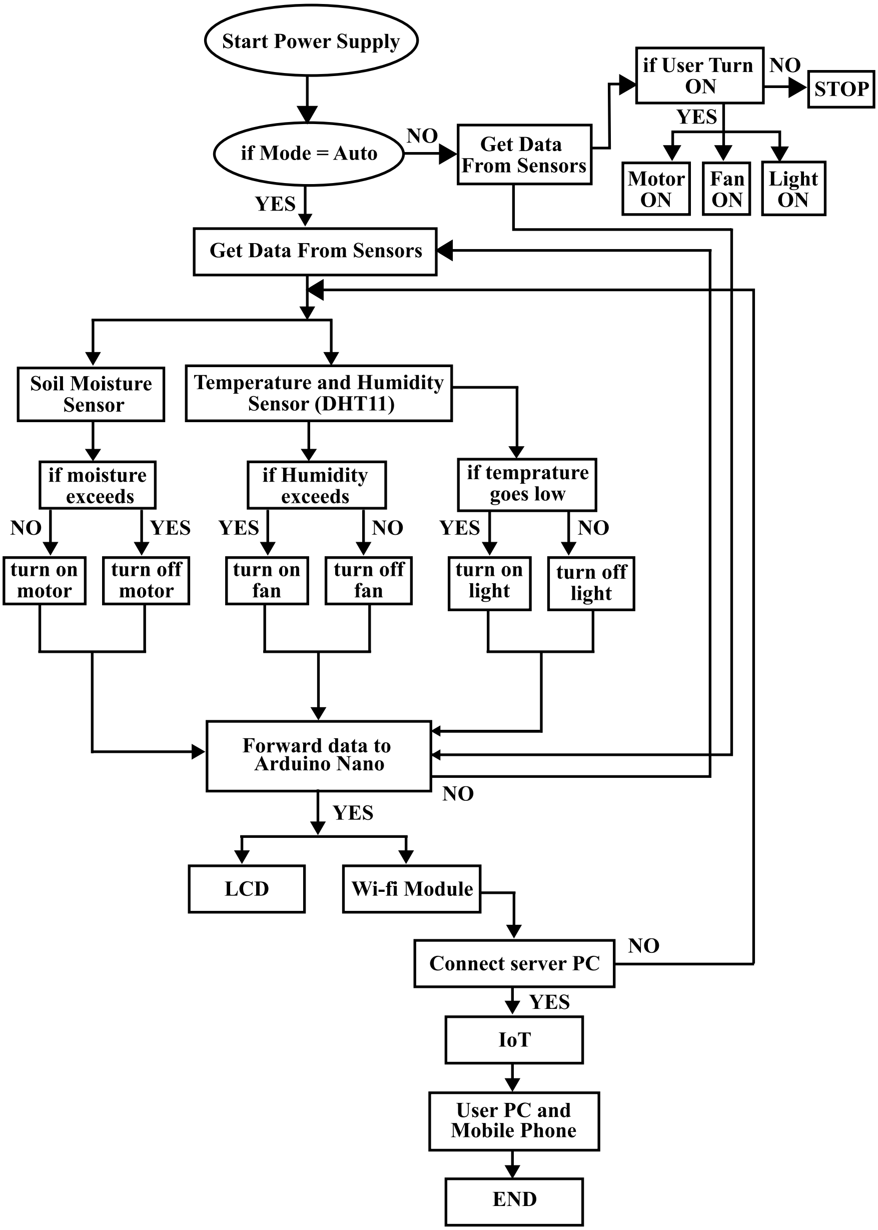


Fig. 6 Flowchart of DIIGMCS

Table 1 Experimental data table

Date and time	Temperature (°C)	Humidity (%)	Soil moisture (%)
17-10-2021 3:23	28	65	40
17-10-2021 3:24	28.5	65	70
17-10-2021 3:25	28.6	65	70
17-10-2021 3:26	28.45	70	90

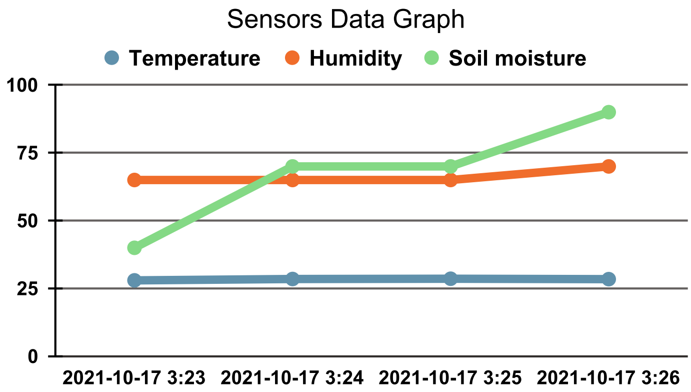


Fig. 7 Sensors data graph of DIIGMCS

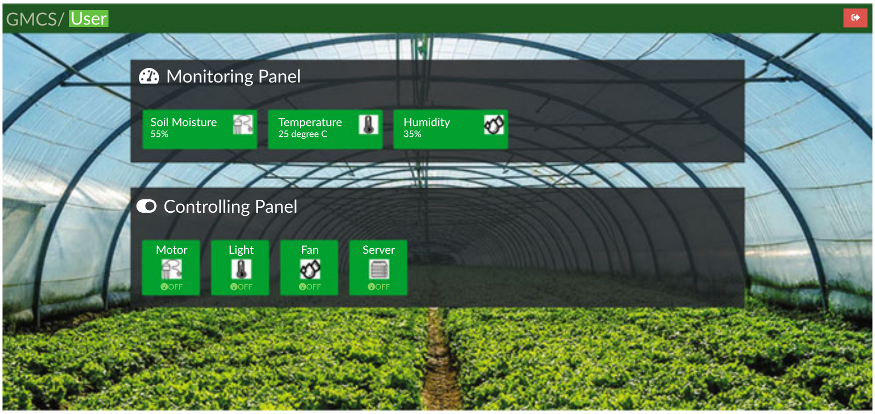


Fig. 8 Sensors output data and relay switches current status showing on IoT Website

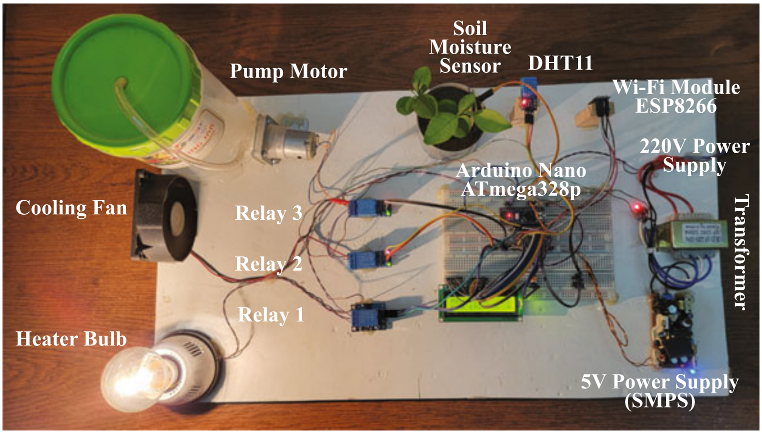


Fig. 9 System prototype of DIIGMCS

6 System Prototype

A prototype is a framework on which something is based that is a sample version of a final product. The system prototype is shown in Fig. 9.

7 Conclusion

In this work, an IoT-based greenhouse system has been designed and effectively proved such as IoT that assures continuous monitoring of several environmental parameters that helps a farmer to monitor and control the parameters even distantly using mobile phone. From this proposed system, it is concluded that automation in wireless sensors plays vital role to design a web-based greenhouse monitoring system to continuously monitor the humidity, temperature, and soil moisture. Moreover, the proposed system is in a place to continue a long-term analysis on greenhouse environmental conditions. The progress in communication engineering, automatization, and availability of inexpensive single-chip microcontrollers has encouraged the analyzer to implement IoT-based greenhouse system.

References

1. Sampaio, H., Motoyama, S.: Implementation of a greenhouse monitoring system using hierarchical wireless sensor network. In: Proceedings of the IEEE 9th Latin-American conference on communications (LATINCOM), Guatemala City (2017)
2. Gomes, T., Brito, J., Abreu, H., Gomes, H., Cabral, J.: GreenMon: an efficient wireless

sensor network monitoring solution for greenhouses. In: Proceedings of the IEEE international conference on industrial technology (ICIT), Seville (2015)
3. Li, L., Yang, S., Wang, L., and Gao, X.: The greenhouse environment monitoring system based on wireless sensor network technology. In: IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, Kunming (2011)
4. Dan, L., Xin, C., Chongwei, H., Liangliang, J. I.: Intelligent agriculture greenhouse environment monitoring system based on IoT technology. In: Proceedings of the international conference on intelligent transportation, big data and smart city, Halong Bay (2015)
5. Vimal, P.V., Shivaprakasha, K.S.: IOT based greenhouse environment monitoring and controlling system using Arduino platform. In: IEEE International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kerala (2017)
6. Yahaya, A., Abass, Y.A., Adeshina, S.A.: Greenhouse monitoring and control system with an Arduino system. In: IEEE 15th International Conference on Electronics, Computer and Computation (ICECCO), Abuja (2019)
7. Liu, L., Zhang, Y.: Design of greenhouse environment monitoring system based on wireless sensor network. In: Proceedings of the IEEE 3rd International Conference on Control, Automation and Robotics (ICCAR), Nagoya (2017)
8. Kodali, R.K., Jain, V., Karagwal, S.: IoT based smart greenhouse. In: Proceedings of the IEEE humanitarian technology conference (R10-HTC), Agra, Dec 2016
9. Yoon, C., Huh, M., Kang, S., Park, J., Lee, C.: Implement smart farm with IoT technology. In: 2018 IEEE 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon (2018)
10. Ahmed, N., De, D., Hussain, I.: Internet of things (IoT) for smart precision agriculture and farming in rural areas. IEEE Internet Things J. **5**(6) (2018)
11. Geethamani, R., Jaganathan, S.: IoT Based smart greenhouse for future using node MCU. In: 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore (2021)
12. Vishwakarma, A., Sahu, A., Sheikh, N., Payasi, P., Rajput, S.K., Srivastava, L.: IOT Based greenhouse monitoring and controlling system. In: IEEE Students Conference on Engineering & Systems (SCES), Prayagraj (2020)
13. Maria, C., Oscar, M., Suarez, J., Garcia, A. P.: IoT-based automated greenhouse for deep water culture hydroponic system. In: 2nd Sustainable Cities Latin America Conference (SCLA), Medellin (2021)
14. Karlo, S., Tolentino, Edmon, O., Fernandez, R., et al.: Development of an IoT-based aquaponics monitoring and correction system with temperature-controlled greenhouse. In: IEEE International SoC Design Conference (ISOCC), Jeju (2019)
15. Kodali, R.K., Sahu, A.: An IoT based soil moisture monitoring on Losant platform. In: IEEE 2nd international conference on contemporary computing and informatics (IC3I), Greater Noida (2016)
16. Ezhilazhahi, A.M., Bhuvaneswari, P.T.V.: IoT enabled plant soil moisture monitoring using wireless sensor networks. In: 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), Chennai (2017)
17. Athani, S., Tejeshwar, C.H., Patil, M., Patil, P., Kulkarni, R.: Soil moisture monitoring using IoT enabled arduino sensors with neural networks for improving soil management for farmers and predict seasonal rainfall for planning future harvest in North Karnataka—India. In: International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam (2017)
18. Du, J., Guo, J., Xu, D., Huang, Q.: A remote monitoring system of temperature and humidity based on OneNet cloud service platform. In: IEEE Electrical Design of Advanced Packaging and Systems Symposium (EDAPS), Haining (2017)
19. https://store.arduino.cc/products/arduino-nano. Last accessed 27 Nov 2021
20. https://create.arduino.cc/projecthub/MisterBotBreak/how-to-use-a-soil-moisture-sensor-ce769b. Last accessed 27 Nov 2021

Remote Measurement of Nitrogen and Leaf Chlorophyll Concentration Using UAV-Based Multispectral Imagery from Rice Crop Field in Sri Lanka

P. P. Dharmaratne, A. S. A. Salgadoe, W. M. U. K. Rathnayake, A. D. A. J. K. Weerasinghe, D. N. Sirisena, and W. M. N. Wanninayaka

Abstract Rice is the staple food in Sri Lanka and the widely cultivated crop among farmers. Proper assessment of crop nitrogen and rice crop condition in on-going cultivation helps farmers to effectively manage the crop inputs. Conventional methods of assessing rice crop condition, i.e., crop greenness by human eyes and leaf color chart method, are subjective and have limitations in converting in to quantitative decisions. Thus, assessing the nitrogen level, leaf greenness and leaf chlorophyll concentration has become challenging with traditional methods in Sri Lankan rice industry. This study evaluates the performance of applying remote sensing technique using unmanned aerial vehicle (UAV) for non-destructive, measurement of crop nitrogen level, leaf greenness and chlorophyll concentration. Multispectral UAV images were acquired using multispectral drone from a controlled rice field (Bg 300, at booting stage) in the rice research station with four blocks of treated nitrogen application levels. On-ground measurements were taken from quadrant size (1 m × 1 m) sample areas in each block. From the UAV-derived RGB orthomossaic and reflectance maps, normalized vegetative index (NDVI) was calculated. Finally, the averaged NDVI values extracted from the quadrant areas of the rice crops were compared against the ground measured values using Pearson correlation fit analysis. The results proved that NDVI strongly correlated with leaf chlorophyll, leaf greenness and nitrogen level with $R^2 = 97\%$, 96.8% and 96.4%, respectively. Findings strongly suggest the possibility of remotely measuring of nitrogen and chlorophyll level of the rice crop field.

Keywords Drone · Multispectral imagery · Nitrogen levels

P. P. Dharmaratne (✉) · A. S. A. Salgadoe
Department of Horticulture and Landscape Gardening, Faculty of Agriculture and Plantation Management, Wayamba University of Sri Lanka, Makandura 60170, Gonawila, Sri Lanka
e-mail: dharmaratnep@gmail.com

W. M. U. K. Rathnayake · D. N. Sirisena · W. M. N. Wanninayaka
Rice Research and Development Institute, Bathalagoda, Ibbagamuwa 60500, Sri Lanka

A. D. A. J. K. Weerasinghe
Retail Information Technologies Pvt. Ltd., Bambalapitiya, Sri Lanka

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_46

1 Introduction

Rice (*Oryza sativa* L.) is one of the most important grain crops in the world, especially in Asian countries. The improvement of crop production to fulfill the increased market for agricultural products brought on by the growing global population is one of the biggest problems of the twenty-first century. More than 1.8 million farmers' livelihood are depend on rice farming in Sri Lanka, and with total population of 21 million, Sri Lanka's per capita annual rice consumption is approximately 107 kg per person. Usually, farmers apply high inputs, i.e., excessive application of nitrogen fertilizer in rice cultivation with the expectation of having a high yield [1]. In order to determine the best time to harvest their crops, enhance agronomic management techniques and enhance crop inspections, among other things, farmers are calling for more and more quick, affordable, environmentally friendly and non-destructive approaches of precision farming [2]. Efficiency in the use of water or nitrogen is one such concept that has been around for many years. This raises production costs and puts the environment at danger. Input application can be effective or target application can be encouraged if farmers have a field-level crop evaluation method to determine the current nitrogen levels in the rice field during on-going cultivation before harvest. Modern agricultural remote sensing techniques are far more advantageous since they offer great potential for quantitatively analyzing the crop without destroying it, unlike existing traditional crop evaluation methods, which have limitations.

1.1 Estimation of Leaf Greenness

Usually, farmers use the rice crop 'greenness' based on the leaf color chart' (LCC) to assess the crop nitrogen level in field [3]. The LCC has a linear relationship with available nitrogen level in soil, where high LCC values indicate high greenness associated with good nitrogen (N) availability in soil [3]. Destructive sampling and the necessity of laboratory testing hinder the practical application of this method by rice farmers. Simple diagnostic tools have been developed to monitor plant nitrogen status and for fine-tuning of nitrogen management, and the LCC and SPAD meter (chlorophyll meter) are the two common methods among those simple diagnostic tools. The main disadvantage of these two tools is that there is no calibration available at local conditions to justify whether the critical nitrogen content of the plant at the time of measurement has been achieved [4]. The leaf N content is closely related to photosynthetic rate; it is a sensitive indicator of the dynamic changes in crop N demand within a growing season [5]. Therefore, measurement of leaf N content at different N levels is a good indicator to determine the N response of rice. The direct measurement of leaf N concentration using the conventional methods is laborious, time-consuming and costly. However, assessing greenness of the crop by humans is subjective and late indicative of crop condition. A study conducted by Sirisena [3] showed a strong relationship between crop nitrogen levels measured using the crop greenness with LCC. However, use of LCC is not practically feasible to conduct in an on-going cultivation by a normal farmer due to the lack of knowledge and difficulty in applying it to the whole field. Therefore, farmers need an accessible technology to assess crop performance in their field before the end of the cropping season. Nitrogen is one of the most yield-limiting nutrients in crop production [6]. Given the strong correlations between leaf nitrogen content and chlorophyll, by obtaining a quick and accurate measurement of the latter, the farmer is able to measure the level of nitrogen in plants.

1.2 Chlorophyll Concentration of Rice Crop

In practice, the chlorophyll content of leaves is often used to predict the physiological condition of the leaves, as influenced by various natural factors. Traditionally, crop assessment had relied on ground-based field survey and visual observation to measure plant status by collecting a small sample size [7].

Chlorophylls are a significant and essential component of the photosynthesis process, and this is appropriate because remote sensing can evaluate plant physiological growth. According to estimates, healthy plants with the ability to demonstrate growth rates till they reach their full potential have more chlorophyll than unhealthy plants [8]. Accordingly, based on its nutritional form, the recognition and tracking of chlorophyll content in a leaf can be used to detect and study plant stress. This has a significant impact on crop condition and circumstances, the degree of seriousness and the necessary nutrients, particularly in precision agriculture practices. The techniques used to harvest chlorophyll from plants almost often rely on damaging techniques. Additionally, the SPAD chlorophyll meter offers a quick and non-destructive method for measuring chlorophyll levels outdoors. These techniques, however, take a lot of time and extra labor. The electromagnetic spectrum's 670 to 780 nm wavelength range is where the spectral absorption typically takes place [9]. In this case, leaf chlorophyll exhibits substantial absorption at 450 and 670 nm and high reflection at 700–900 nm in the near infrared.

1.3 Crop Remote Sensing for Rice Cultivation

Crop monitoring has traditionally relied on fieldwork-based data collection methods.

Such a method results in inaccurate crop assessment and crop area estimation because it is frequently subjective, expensive and prone to significant errors [10]. Additionally, it can take some time for the data to be made public so that proper steps can be done to prevent a food crisis. However, unless created or proven locally, no sufficient crop monitoring techniques are suitable for usage locally [11]. Remote sensing (RS) techniques have the ability to quantitatively, instantly and nondestructively give information on agricultural crops over vast distances. Remote sensing, in general, refers to the process of gathering data about an object or geographic phenomenon without physically touching it and evaluating it [11]. Monitoring the rice field to understand the growing state is necessary since rice crops are crucial. Conditions of agricultural crops, such as rice, may now be quickly monitored from the air thanks to advancements in technology. The procedures can be viewed using a camera or sensor mounted on a helicopter, an airplane, a drone or a satellite [10, 12].

Visual observations and the fundamentals of remote sensing using satellites, unmanned aerial vehicles (UAV) and other technologies are comparable. The green color of plants is easily recognized by human sight as reflected energy that bounces off their leaves. The energy for heat or photosynthesis is provided by wavelengths that are absorbed by the plants [13]. Since it is better for photosynthesis, green reflected light more than blue or red light. A plant appears green because its leaves contain chlorophyll, which absorbs a large amount of light in the visible spectrum and reflects the green hue [14]. Crop monitoring has traditionally involved applying various direct and indirect measures of plant features, soil parameters and environmental conditions [15]. The most accurate monitoring strategy is ground-based measurements. However, the method is expensive and necessitates a lot of human resources. Earth observation (EO) satellite tools and the related datasets have been widely used for decades as alternative means for agriculture crop monitoring over a broader geographical area (e.g., MODIS, Landsat, Sentinel, etc. [16, 17]. Cloud cover is a problem for satellite remote sensing since it reduces the resolution of the pixels in the hazy imagery. Compared to satellite photos, UAVs are emerging as new, viable platforms for collecting spatial data at a lower cost [18].

Contrarily, UAV technology's superior spatial resolution allows it to distinguish finer elements of a region, and this form of imaging offers fresh approaches to crop management and field monitoring [18].

Recent UAV technology offers the ability to boost picture data monitoring frequency and geographical resolution. For collecting the distinct reflectance signal of plants in a number of short wavelengths of the visible, red edge, or NIR areas, a lightweight multispectral camera is typically created [19]. Chlorophyll fluorescence calculations made from UAV-captured narrow-band multispectral pictures have also intermittently been used to address the issue of identifying and monitoring water stress. The ability to fly at low altitudes easily, ultra-high spatial resolution images (i.e., centimeter-level) and on-going navigation and flight control system development and enhancement all present excellent data opportunities for monitoring the crop planning process at the local level [20].

2 Normalized Difference Vegetation Index (NDVI)

NDVI calculations based on UAV-based photos have established themselves as a reliable technique for crop monitoring. Based on the quantity of chlorophyll content, vegetation indices are frequently used to estimate crop status utilizing the visible and near-infrared (NIR) sections of the electromagnetic spectrum [21]. Chlorophylls have strong red-region characteristic absorption and significant near-infrared-region reflectance peaks [22]. Between 660 and 680 nm, the red region's maximum absorption occurs. This is because, with the exception of high chlorophyll concentration, the absorption range of 660–680 nm tends to be overwhelmed at low chlorophyll levels and rebound in the near–infrared region, decreasing the sensibility of the spectral indices based on this wavelength [23]. We used a multispectral UAV platform to quickly track the NDVI.

The rate of leaf expansion is constrained when there is a nitrogen deficiency. Because of this, spectral profile variations found by numerous analyses may be more closely related to variations in the ground cover than to the actual spectral characteristics of the plant. High-resolution photos are frequently needed to address potential problems, limit the amount of mixed pixels and correctly resolve troublesome areas. By thresholding the NDVI image with related NDVI values, the undesirable soil and background pixels can be eliminated [24]. This research impacted the use of UAVs to monitor the nutritional health of the rice crop using indices developed from canopy reflectance spectra obtained at the booting stage. However, the foundation for making precise managerial decisions is a map of the vegetation index (VI). Farmers may easily monitor crop development and the status of the paddy in real-time by having a plot of the vegetation indices.

2.1 Purpose of the Paper

Farmers who are assessing the crop greenness, visually are lack with accessible technology to quantify the variation of crop nitrogen level in an on-going rice field. Greenness of the crop is subjective and late indicative of crop condition for human eyes. So, farmers' existing crop management practices are based on qualitative judgments rather than quantitatively measured parameters of crop nitrogen level status. Due to this reason at the end of the cultivation season farmers face challenges with high inputs and costs of production but resulting low yield.

Therefore, this study was designed with the aim of monitoring chlorophyll concentration of leaves, leaf greenness and crop nitrogen level, developing using a UAV-based remote sensing protocol, so that the farmers and researches can monitor the crop response during management activities real-time before the harvest (Fig. 1).

3 Materials and Methods

3.1 Field Experimental Design

The initial calibration study with UAV surveying was carried out at the rice fields maintained by the Rice Research and Development Institute, Bathalagoda, Sri Lanka.

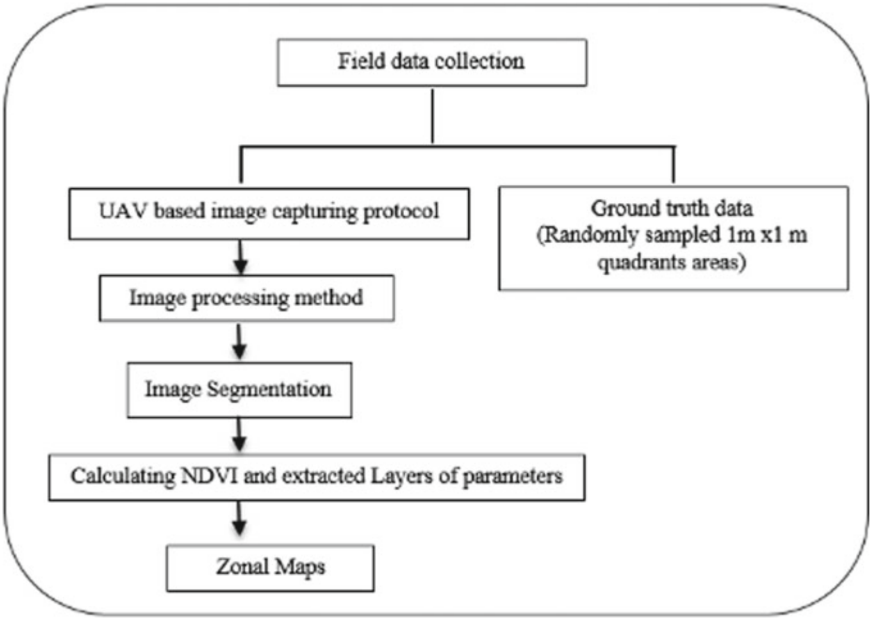


Fig. 1 Method of the study

Rice seeds (Bg 300 variety) were sown in middle March, transplanted in early April and harvested in July in 2021. The experiment was conducted in monitoring the crop nitrogen level, leaf greenness and leaf chlorophyll concentration, and for this study, complete randomize design was used. There were four treatments as 0%, 50%, 100%, and 150% of nitrogen fertilizer levels. And, 1m^2 white regiform quadrants were laid as the sampling units (Fig. 2).

3.2 On-Ground Data Collection

The ground-based measurement was carried out after flight of survey, at the booting stage. Concentration of chlorophyll in leaves was taken using a SPAD-502 chlorophyll meter, The ten fully expanded youngest leaves were randomly selected from each sampling unit/quadrant and measured [3]. The leaf greenness was measured using the leaf color chart introduced by the RRDI. Ten plants from each quadrant were selected, and the middle portion of the fully expanded youngest leaf was measured. On-ground measurements, i.e., leaf chlorophyll and leaf color chart greenness, were taken from quadrant size (1 m × 1 m) sample areas (demarcated by white regiform frame tags) in each block.

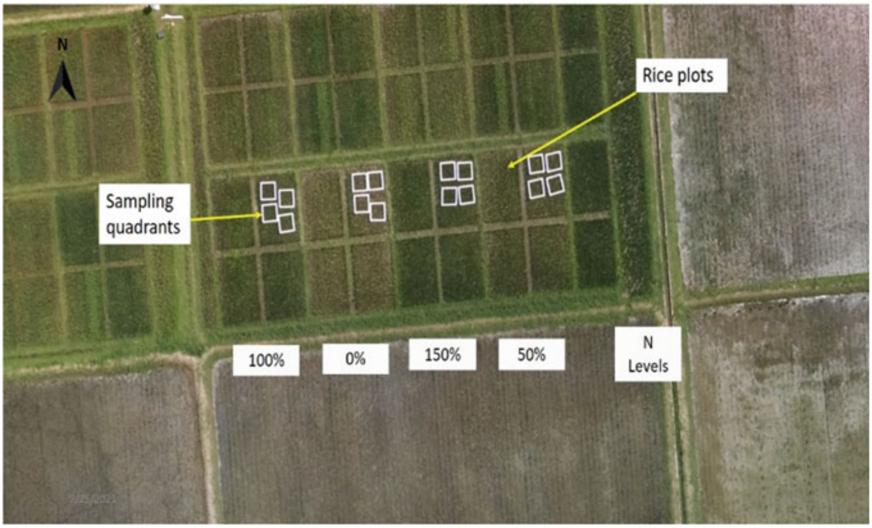


Fig. 2 Orthomosaic RGB image captured and generated by Pix4D mapper, at 1½ month before the harvest

3.3 UAV-Based Data Collection

3.3.1 Image Acquisition by UAV

All imagery data was captured using the P4 Multispectral (DJI Technology) drone camera with sensors for wavebands in blue: 465–485 nm, green: 550–570 nm, red: 663–673 nm, red edge: 712–722 nm and near-infrared: 820–860 nm (Fig. 3). Physical radiometric targets were imaged prior to flight for radiometric calibration. Mission planning was conducted using DJI GS Pro app, and the images was improved into Pix4Dmapper for processing. Field trial was surveyed from the height of 45 m. Flying was conducted between 08:00 h and 10:00 h. The flight speed was 2 ms^{-1}, and images was captured parallel to the path, resulting in front and side overlaps of approximately 80%.

3.3.2 Image Processing

The multispectral images were processed to orthomosaic and produced a single raster image file using the Pix4D mapper software (Fig. 2).

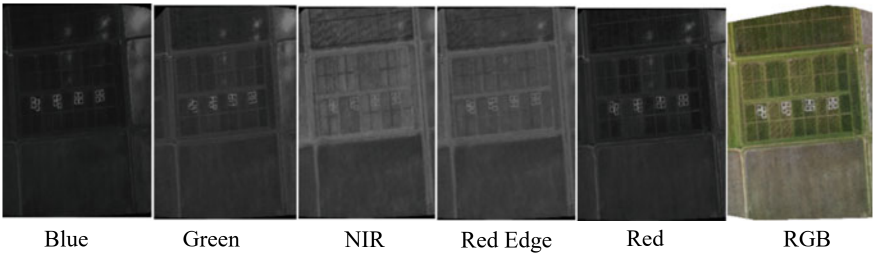


Fig. 3 Orthomosaic maps generated by Pix4D mapper. Left: the reflectance maps per channel captured by the camera model FC6360, right: the RGB map capture by the Camera model FC6360

3.3.3 Image Segmentation

Segmentation targeted rice plants from the background were conducted using the NDVI as the threshold. The NDVI related to the rice plants was manually identified and threshold it from the ground NDVI. Normally the ground (soil, water) has a lower NDVI than the vegetation. The image classification was done using QGIS version 3.20.

3.3.4 Calculation of Vegetation Indices

NDVI was computed from the processed drone image to compare with the ground-truth data. VIs and the raw wavebands were used as remotely retrieved measurement of the rice plant. From the UAV-derived RGB orthomossaic and reflectance maps: red, green, blue, red-edge and near infrared (NIR), raster images normalized vegetative index (NDVI) was calculated. The undesired soil and background pixels were removed by thresholding the NDVI image with associated NDVI values, respectively (Fig. 1).

3.4 Data Analysis

The averaged NDVI values extracted from the sampled quadrant areas (identified by the visible white-color regiform tags) of the rice crops were compared against the ground measured values using Pearson correlation fit analysis. The data analysis which was conducted with the Pearson correlation best-fit analysis where vegetation index (NDVI) with the ground-truth physiological parameters were compared. Furthermore, derived normalized difference vegetation index (NDVI) values from the drone imagery in the sample rectangular boundary areas were analyzed with the ground-truth data, i.e., chlorophyll content and leaf greenness to establish a co-relation fit.

4 Results

4.1 Determination of Field-Measured Agronomic Traits and NDVI

The results suggested that NDVI and the concentration of chlorophyll in leaves were achieved with the R^2, 0.97 (Fig. 4a). As shown in Fig. 4b the leaf greenness also conducted a comparable NDVI with the R^2, 0.97. Also, the nitrogen level was best fitted with NDVI with the R^2, 0.96 (Fig. 4c and Table 1).

Table 1 Readings of concentration of chlorophyll (SPAD readings) in leaves and NDVI according to the N levels

N Level (%)	Chlorophyll concentration of leaves	NDVI (paddy)
0	34.51	0.269
50	37.11	0.443
100	37.54	0.512
150	39.59	0.576

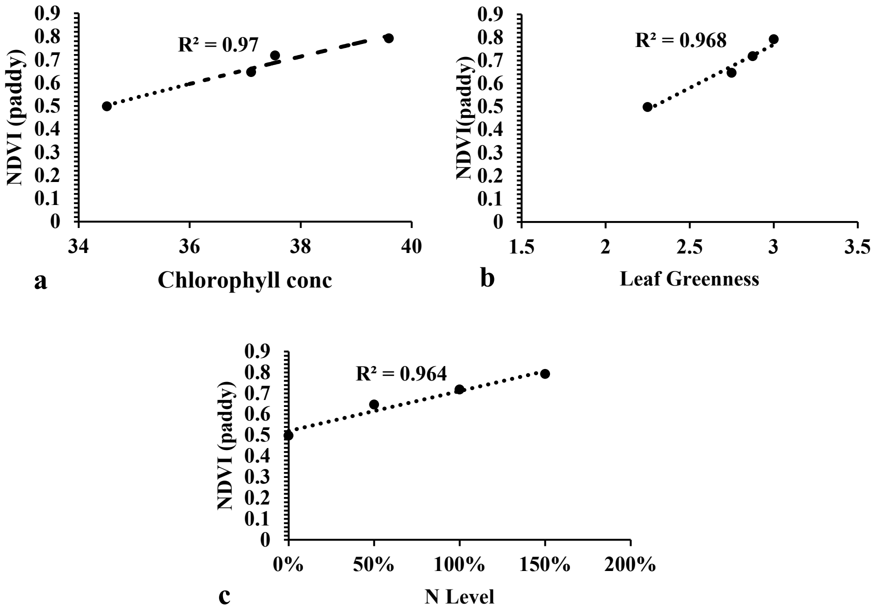


Fig. 4 R^2 value of the Pearson correlation, A-NDVI versus concentration of chlorophyll in leaves, B- NDVI versus leaf greenness and C- NDVI versus nitrogen level

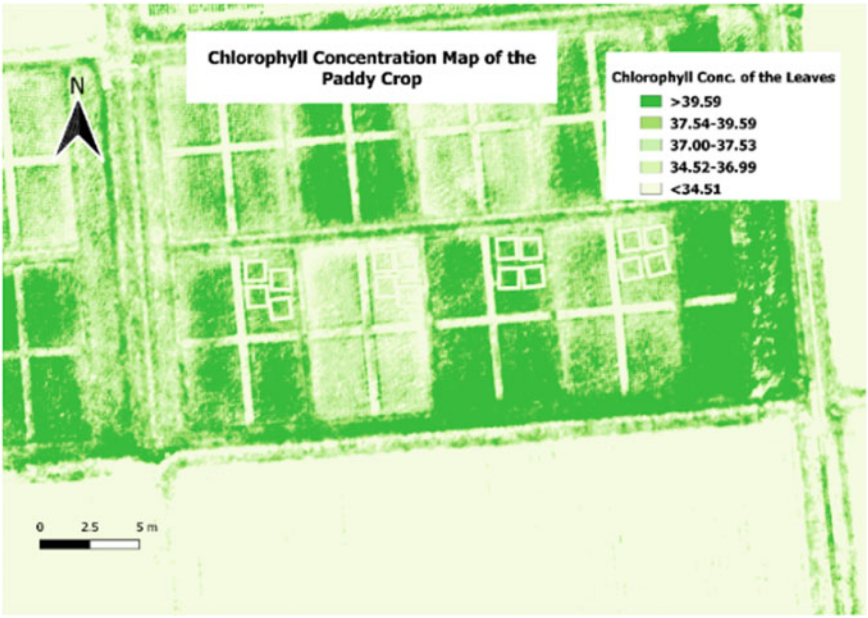


Fig. 5 Chlorophyll concentration map of the paddy crop field at the 45 m (booting stage) analyzed with the NDVI value

4.2 Zonal Map

An interactive chlorophyll concentration of the rice crop and nitrogen level maps (Figs. 5 and 6, respectively) were mapped using QGIS version 3.20 in color print form or in indicating the leaf chlorophyll concentration and the distribution of the nitrogen level. As shown in Fig. 5, the relationship between the NDVI values and leaf chlorophyll concentration showed a significant variation.

N level of 0–150% showed a distribution from dark green to light green, as the SPAD values also range from 39.59 to 34.51. In recommended nitrogen level, it was mostly shown the SPAD value 39 but the distribution is not evenly dark green, only the middle area was 39, border area was 37 like (Fig. 5). As shown in Fig. 6, the nitrogen level was shown a significant distribution with the NDVI value. From the nitrogen level of 0–150%, there was a variation. So, the tool will be applicable for the sustainable farming and farmers can be produced their input and apply the management practices according to the requirements and most important factor is that the farmer knows their crop performance before the harvest and they can attend to their patches with the zonal maps which they receive. These findings highly encourage the use of remote sensing tools in rice cultivation and greater adaptation of UAV-based crop assessment tools in future agriculture.

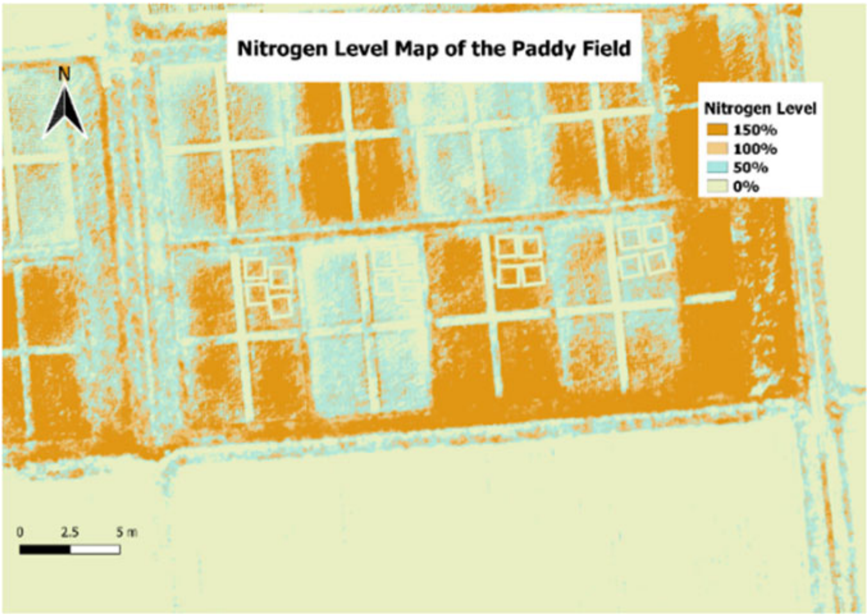


Fig. 6 Nitrogen-level map of the paddy field at the 45 m (booting stage) analyzed with the NDVI values

5 Discussion

It has been suggested that chlorophyll contents are directly connected with crop development because chlorophyll transforms solar energy into chemical energy. To address the demand for precision nitrogen management, it is crucial to create highly effective, trustworthy, and useful systems for monitoring crop nitrogen status [25].

Several conventional methods, such as the leaf color chart [4] or destructive chemical analysis [26], are limited by low efficiency and small-scale applicability.

Crop N-status monitoring and management based on the spectrum have become widely employed in a variety of crops thanks to advancements in optical components and crop remotely sensor based [16]. This study demonstrated that rice crop fields may accurately be assessed for leaf greenness and chlorophyll content using UAV-based multispectral imaging [27]. Additionally, this approach might be simply applied to UAV-based multispectral imaging. As a result, this approach is practical and provides technological support for future grain yield and quality predictions as well as N diagnosis and control. The NDVI and SPAD values had a positive linear connection (Table 1). NDVI and SPAD measurements correlate more strongly. This association is consistent with other research' findings that the vegetation index (NDVI) and ground information like SPAD chlorophyll measurements and nutrient content have a positive and linear relationship. Zhou [28] suggested that vegetation index, such as normalized difference vegetation index (NDVI), could be used to identify rice crops' characteristics.

6 Conclusion

The results of this experiment clearly support the idea that agricultural remote sensing technology can be utilized in rice fields to replace more conventional subjective crop monitoring techniques. With great precision, leaf chlorophyll content and nitrogen level from the rice field could be measured using UAV-based multispectral imaging. The UAV-based methodology can be employed to perform target fertilizer use efficiency or variable fertilizer application, enabling more environmentally sustainable agriculture, in place of on-ground crop varieties assessment methods. The results support the adoption of smart plant remote sensing methods in future rice farming with precise production projections from local, regional, and national levels.

References

1. Wan, L., Cen, H., Zhu, J., Zhang, J., Zhu, Y., Sun, D.: Agricultural and Forest Meteorology Grain yield prediction of rice using multi-temporal UAV-based RGB and multispectral images and model transfer—a case study of small farmlands in the South of China. Agric. For. Meteorol. **291**, 108096 (2020). https://doi.org/10.1016/j.agrformet.2020.108096
2. Congalton, R.G., Gu, J., Yadav, K., Thenkabail, P., Ozdogan, M.: Global land cover mapping: a review and uncertainty analysis 12070–12093 (2014). https://doi.org/10.3390/rs61212070
3. Sirisena, D.N., Wikramasingha, W.M.A.B.D., Ranathunga, R.A.N.D.: Use of leaf colour charts to determine nitrogen requirement and yield response of rice variety. Annu. Sri Lanka Dep. Agric. **8**, 229–238 (2006)
4. Sirisena, D.N., Sumanapala, K.V.K.K., Wanninayake, W.M.N.: Use of leaf colour chart and SPAD meter to estimate total Nitrogen conetent in rice (*Oryza sativa* L.,) Plants. Annu. Sri Lanka Dep. Agric. **14**, 313–314 (2012)
5. Hernández-clemente, R., Navarro-cerrillo, R.M., Zarco-tejada, P.J.: Remote sensing of environment carotenoid content estimation in a heterogeneous conifer forest using narrow-band indices and PROSPECT + DART simulations (a). Remote Sens. Environ. **127**, 298–315 (2012).https://doi.org/10.1016/j.rse.2012.09.014
6. Cui, D., Li, M., Zhang, Q.: Development of an optical sensor for crop leaf chlorophyll content detection. **69**, 171–176 (2009). https://doi.org/10.1016/j.compag.2009.08.0017.
7. Sims, D.A., Gamon, J.A.: Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages **4257** (2017). https://doi.org/10.1016/S0034-4257(02)00010-X
8. García Cárdenas, D.A., Ramón Valencia, J.A., Alzate Velásquez, D.F., Palacios Gonzalez, J.R.: Dynamics of the Indices NDVI and GNDVI in a rice growing in its reproduction phase from multi-spectral aerial images taken by Drones. In: Advances in Intelligent Systems and Computing, pp. 106–119. Springer Verlag (2019)
9. Escalante, H.J., Rodríguez-Sánchez, S., Jiménez-Lizárraga, M., Morales- Reyes, A., De La Calleja, J., Vazquez, R.: Barley yield and fertilization analysis from UAV imagery: a deep

learning approach Barley yield and fertilization analysis from UAV imagery **425** (2019). https://doi.org/10.1080/01431161.2019.1577571
10. Wang, Y.P., Chen, S.H., Chang, K.W., Shen, Y.: Identifying and characterizing yield limiting factors in paddy rice using remote sensing yield maps. Precis. Agric. **13**, 553–567 (2012). https://doi.org/10.1007/s11119-012-9266-5
11. Tahir, M.N., Zaigham, S., Naqvi, A., Lan, Y., Zhang, Y., Wang, Y., Afzal, M., Jehanzeb, M., Cheema, M., Amir, S.: Real time estimation of chlorophyll content based on vegetation indices derived from multispectral UAV in the kinnow orchard, **1**, 24–32 (2018). https://doi.org/10.33440/j.ijpaa.20180101.0001
12. Pettorelli, N., Vik, J.O., Mysterud, A., Gaillard, J., Tucker, C.J., Stenseth, N.C.: Using the satellite-derived NDVI to assess ecological responses to environmental change. https://doi.org/10.1016/j.tree.2005.05.011
13. Chang, K.W., Shen, Y., Lo, J.C.: Predicting rice yield using canopy reflectance measured at booting stage. Agron. J. **97**, 872–878 (2005). https://doi.org/10.2134/agronj2004.0162
14. Cen, H., Wan, L., Zhu, J., Li, Y., Li, X., Zhu, Y., Weng, H.: Dynamic monitoring of biomass of rice under different nitrogen treatments using a lightweight UAV with dual image - frame snapshot cameras. Plant Methods, 1–16 (2019). https://doi.org/10.1186/s13007-019-0418-8
15. Dammalage, T.L., Sirisena, P.M.T.S., Susaki, J.: MODIS satellite data based rice yield forecasting model for Sri Lanka: a pilot study in Kurunegala District, **17**(3), 25–40 (2017)
16. Zarco-tejada, P.J., Guillén-climent, M.L., Hernández-clemente, R., Catalina, A.: Author's personal copy agricultural and forest meteorology estimating leaf carotenoid content in vineyards using high resolution hyperspectral imagery acquired from an unmanned aerial vehicle (UAV) Author's personal copy **172**, 281–294 (2013)
17. Hashemi, S.A., Chenani, S.K.: Investigation of NDVI index in relation to chlorophyll content change and phenological event, 22–28 (2010)
18. Garcia, J., Barbedo, A.: A Review on the Use of unmanned aerial vehicles and imaging sensors for monitoring and assessing plant stresses (2019). https://doi.org/10.3390/drones3020040
19. Verlag, F., Gitelson, A.A., Gritz, Y., Merzlyak, M.N.: Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves, **282** (2003)
20. Xiao, X., Boles, S., Liu, J., Zhuang, D., Frolking, S.: Mapping paddy rice agriculture in Southern China using multi-temporal MODIOS images mapping paddy rice agriculture in southern China using multi-temporal MODIS Images (2005). https://doi.org/10.1016/j.rse.2004.12.009
21. Bernardis, C. De, Vicente-guijalba, F., Martinez-marin, T., Lopez-sanchez, J.M.: Particle filter approach for real-time estimation of crop phenological states using time series of NDVI images, 5–10 (2016). https://doi.org/10.3390/rs8070610
22. Yamada, N., Fujimura, S.: Nondestructive measurement of chlorophyll pigment content in plant leaves from three-color reflectance and transmittance. Applied Optics. **30**, 3964–3973 (1991)
23. Haghighattalab, A., Crain, J., Mondal, S., Rutkoski, J., Singh, R.P.: Application of geographically weighted regression to improve grain yield prediction from unmanned aerial system imagery (2017). https://doi.org/10.2135/cropsci2016.12.1016
24. Kastens, J.H., Kastens, T.L., Kastens, D.L.A., Price, K.P., Martinko, E.A., Lee, R.: Image masking for crop yield forecasting using AVHRR NDVI time series imagery B. **99**, 341–356 (2005). https://doi.org/10.1016/j.rse.2005.09.010
25. Huang, S., Miao, Y., Zhao, G., Yuan, F., Tan, C., Yu, W., Gnyp, M.L., Lenz-wiedemann, V.I.S., Rascher, U., Bareth, G., International, Y., Yang, Z., Inoue, Y., Zhu, Y., Cao, W., Thenkabail, P.S.: Remote Sens. 10646–10667 (2015). https://doi.org/10.3390/rs70810646
26. Asner, G.P., Martin, R.E.: Remote sensing of environment spectral and chemical analysis of tropical forests: scaling from leaf to canopy levels **112**, 3958–3970 (2008). https://doi.org/10.1016/j.rse.2008.07.003

27. Gitelson, A.A., Merzlyak, M.N.: Signature analysis of leaf reflectance spectra: algorithm development for remote sensing of chlorophyll. J. Plant Physiol. **148**, 494–500 (1996). https://doi.org/10.1016/S0176-1617(96)80284-7
28. Zhou, C., Ye, H., Xu, Z., Hu, J., Shi, X., Hua, S.: Estimating maize-leaf coverage in field conditions by Applying a machine learning algorithm to UAV remote sensing images. Applied sciences.**9**, 1–18 (2019). https://doi.org/10.3390/app9112389

A Method and Design for Developing an IoT-Based Auto-Sanitization System Powered by Sustainable Earth-Battery

Md. Sayeduzzaman, Md. Samiul Islam Borno, Khadija Yeasmin Fariya, and Md. Tamim Ahmed Khan

Abstract Life in the modern world and technology are inextricably linked. The present era is known as the "Age of Science" because of scientific advancements. The Internet of Things (IoT) is a network of interconnected computing devices, mechanical and digital machinery, and items transporting data without requiring human-to-human or human-to-computer contact. An IoT-based automated sanitization system could be designed to stop the spread of COVID-19, the current global pandemic. In this modern era, combining IoT with alternative power sources to produce something positive for people would dramatically accelerate society. Combining an IoT-based sanitization system with an Earth-Battery, the entire prototype can be powered continuously without relying on an external power source. Earth-Battery is a renewable energy source where the soil cells generate electricity in the presence of water. As a result, it's also known as a water-activated battery. A pre-programmed NodeMCU will control disinfector machines and automated doors in a prototype automated sanitization system. The fully automated process can be accessed globally using a smart device and the Internet. This research aims to develop an automated sanitization system that might be applied to safeguard human health by cleaning the interior surface of the atmosphere by getting power from the Sustainable Earth-Battery. A creative solution and incredible potential are designed in this paper to address the energy crisis in society and, with the low installation cost, ensure safety against COVID-19.

Keywords IoT · COVID-19 · Automation · Auto-sanitization system · Disinfector machine · Sustainable Earth-Battery · Renewable power · Health and safety

Md. Sayeduzzaman (✉) · Md. Samiul Islam Borno · K. Yeasmin Fariya · Md. Tamim Ahmed Khan
American International University, 408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
e-mail: sayedz.shuvo@gmail.com

Md. Samiul Islam Borno
e-mail: samiul.borno@gmail.com

Md. Sayeduzzaman · Md. Samiul Islam Borno
Bicchuron-19 Start-Up Project, Team Unknown Inventors, 567/1, Road-5, Sheroil Kolony, Ghoramara, Boalia, Rajshahi 6100, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_47

1 Introduction

After the December 2019 outbreak in China, the World Health Organization (WHO) identified SARS-CoV-2 as a new type of Coronavirus in early 2020. The disease quickly spread globally. On March 11, 2020, the WHO declared COVID-19 a pandemic. COVID-19 has infected more than 26.5 million people in 219 countries. Physicians assume COVID-19 is a disease caused by the SARS-CoV-2 virus that causes respiratory infections. COVID-19's growth has wreaked havoc on human life and caused many deaths globally. According to the worldometer, 234,608,466 people have been affected, with 4,798,466 lives lost, and that number is rising. As a result, many countries had to change their lifestyle [1]. Several international organizations have developed vaccines to combat this deadly virus.

Vaccines don't always work 100%. So, an IoT-based automated sanitization system might be the best option for reducing community transmission, as it does not require human contact. Auto-sanitization uses evaporated fog to clean the air. Antiseptic fluids evaporate at low levels over time to disinfect living spaces, offices, and other surfaces. It can help reduce the COVID-19 epidemic by providing a healthy indoor environment at a low cost. Also, IoT devices use less energy and can be controlled from anywhere. The Internet of Things-based auto-sanitization technology will reduce pandemic impacts while protecting users' health and safety. Due to the rapid advancement of wireless technology, this study has encouraged the authors to operate gadgets via smartphones, one of today's most widely used technological devices. The Internet's importance increases this little device's value. IoT is used in this prototype to reduce human labor while increasing comfort and efficiency. It was designed to keep public health and safety in mind when disinfecting surfaces. A renewable Earth-Battery powered the sanitization system. Using soil, water, and electrodes as raw materials makes it a clean energy source like renewable energy concepts in environmental research. Electricity demand is outpacing supply nowadays. Electricity is vital to human progress. Our country's electricity is generated mainly through natural gas and coal. Neither these resources nor CO_2 emissions are renewable. So, as electricity demand grows, Sustainable Earth-Batteries may become a viable power source. It generates cheap and easy-to-use power from the soil. It is a small battery that can be placed almost anywhere.

For this reason, it is unlikely that any fire will occur in the residential area. No CO_2 is emitted, making it eco-friendly. The Earth-Battery can be fully recharged in 2–3 hour without external sources. It aims to improve everyone's health without polluting the environment. This study shows how to automate such a healthy environmental system using a renewable Earth-Battery system.

The remaining seven parts have been chronologically organized. Section 2 represents the overall literature review of the study, while Sect. 3 represents the methodology and modeling of a sustainable Earth-Battery powered automated sanitization system. Section 5 explores the hardware and software design of the work in progress, while Sect. 6 focuses on the complete working procedure. The future's scope and application are defined in Sect. 7. The research results were briefly discussed in Sect. 8. Finally, Sect. 9 summarizes the findings of this research in the conclusion.

2 Literature Review

In this pandemic, a smart IoT-based automated sanitization system powered by an Earth-Battery might be the most efficient way to reduce costs and fight against COVID-19. There aren't many research subjects relevant to this entire study. However, many previous types of research apply to both portions of this study.

Researchers have proposed some ideas for an automated sanitization system. In this circumstance, a smart IoT-based automated sanitization system might be the most effective way to tackle the corona pandemic. Several researchers have proposed a few possible answers. Sayeduzzaman M., Borno M. I. published a paper on a smart home with an auto-sanitization system, an innovative way to eliminate germs and diseases in a straightforward step. Auto-sanitization is a cleaning method that uses an antiseptic fluid that evaporates on a superficial level over a certain period to disinfect a home's interior surface [2]. One research project involved using ultraviolet light to automatically sanitize items and artificial intelligence. The study focused on cleaning objects' surfaces [3]. The study's main flaw is the inability of U.V. radiation to sterilize the human body. Also, researchers have devised an automated sanitization method using IoT-based robots. It can spray gas and has an optical camera to monitor the spraying activity [4]. In 2019, Nnebedum et al. [5] proposed the Earth's Earth-Battery using magnesium anode and coke cathode combinations. Earth-Battery was published by Belov et al. [6] in 2018. In 2013 Pakpoom Chansri et al. proposed a Study of Marl Soil Potential in Electricity Generation [7]. Borno, Md. Samiul Islam et al. [8] proposed a Journal Article based on the Sustainable Earth-Battery, which generates electricity from the soil where the individual Earth cells contain soil, water, and electrodes such as Carbon (C_6) and Aluminum (Al_{13}). They have also published an article in the book Hybrid Intelligent Systems titled "A Convenient Method and Design for Constructing an IoT-based Smart Automated Sanitization System." In this article, they explain how those of us who appreciate researching with diagrams and images can make the prototype on our own [9]. They came up with the idea of developing an IoT-based automated sanitization system to mitigate the deadly Corona virus. That uses green and alternative energy sources like Sustainable Earth-Battery, the article's primary focus. An IoT-based automated sanitation system can be used indoors or outdoors, depending on the location. A disinfector door detects human movement and immediately begins sanitizing and measuring the individual's body temperature. The door will then automatically open. Sustainable Earth-Battery will power the prototype's entire process.

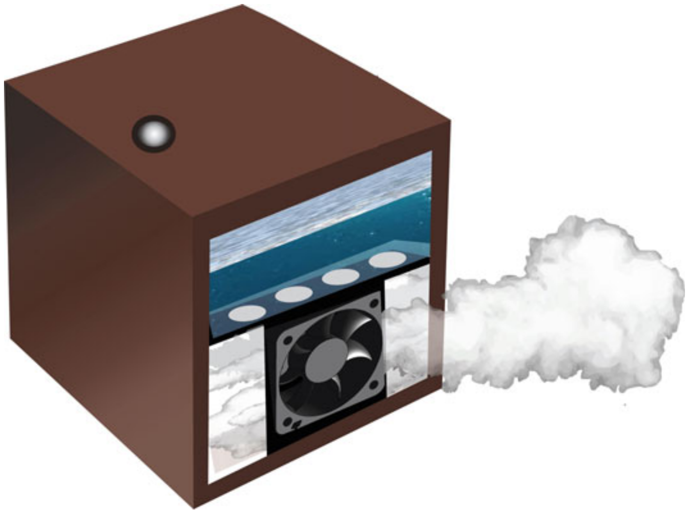


Fig. 1. 3D model of the disinfector machine

3 Methodology and Modeling

3.1 Automated Sanitization System

An automated sanitization system is a virus-cleaning technique that utilizes an antiseptic solution that evaporates to disinfect the surface. The disinfection fog machine kills molds and bacteria by gradually separating their synthetic structure and isolating them from oxygen before expanding. A disinfector machine utilizes an ultrasonic mist generator to evaporate fog to disinfect the environment where it is placed for a predetermined coded time (Fig. 1).

3.2 Sustainable Earth-Battery Prototype

A Sustainable Earth-Battery will power the prototype. This green energy battery is made of soil and water using a confined sphere. Soil's potential energy has been converted into electrical energy. So this massive electron cloud remains underground. An artificial electromagnetic field (AEMF) is used to generate electricity from the earth. The initial force generated by anode and cathode dipole electrodes carbon was used as an anode and aluminum as a cathode in a container. Anode attracts electrons that cathode repels. The series element connects the anode and cathode using capacitive soil, and the Earth-Battery's potential energy has been transformed into electrical power. As the battery activated by water, it is also known as a water

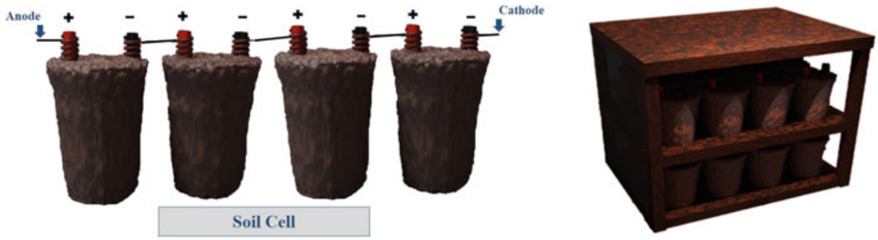


Fig. 2 Soil cells of Earth-Battery and 3D model of the Sustainable Earth-Battery prototype

activated battery. Each Earth-Battery cell produces approximately 1 V DC Soil cells, and a 3D model of the Sustainable Earth-Battery cell configurations of prototype had shown in Fig. 2.

3.3 Automated Sanitization System with the Supply of Power from Sustainable Earth-Battery

The automated sanitization system and a Sustainable Earth-Battery are merged with NodeMCU to make this prototype automated. With the help of IoT devices, i.e., NodeMCU, any smart device can connect via the Internet, an open-source IoT platform at a minimal cost. The NodeMCU firmware is a free and open-source platform with a built-in Wi-Fi integrated circuit [10]. This prototype can run in three ways—Automatically, Remotely, and with the help of using voice commands via google assistant. An Internet connected smartphone is required to enable or disable the disinfector machine and automated door via remote and voice access. Sustainable Earth-Battery had used here to provide continuous power to the sanitization system through a charging circuit that charges the secondary battery from the Sustainable Earth-Battery. The Sustainable Earth-Battery charged secondary batteries could power any 5 V D.C. device. A 3D model of the Sustainable Earth-Battery powered IoT-based sanitization system had shown in Fig. 3 (Fig. 4).

4 Working Procedure

Each disinfector machine contains four ultrasonic 113 kHz mist or fog makers that vibrate at a high frequency and heat the antiseptic liquids to create fogs. The mist or fog creators run at 5 V DC. A brushless fan operating at 5 V D.C can evaporate the mist into the atmosphere. It works by vaporizing the entire inner or outer atmosphere to sanitize the surface atmosphere. A disinfector machine uses non-alcoholic, non-flammable pure antibacterial fluids. Thus, the fog spreads throughout the house, cleaning the inside surface of the living space. A disinfector fog machine sanitizes

Fig. 3 IoT-based auto sanitization system powered by Sustainable Earth-Battery

Fig. 4 A visual representation of a smart IoT-based automated sanitization system powered by a Sustainable Earth-Battery installed in a home

with hydrogen disinfectant, hydrogen peroxide (Food Grade), or a sterile material. Figure 5 illustrates a disinfector machine and its operation.

The utilization of a sustainable Earth-Battery and an automated sanitation system is linked. The eco-friendly Earth-Battery continuously gives power to the sanitization system. The standard sustainable Earth-Battery has 6 V, 100 mA produces from 25 soil cells, four branch connections, and six soil cells arranged in each branch. One cell has been used as a backup to maintain the output stable voltage, and no external I/P power is required to charge the Earth-Battery. The secondary battery can be

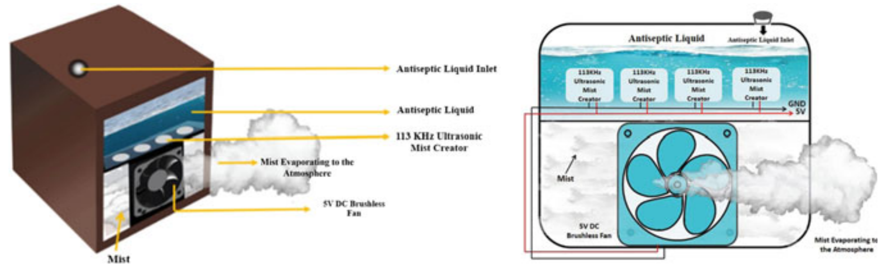


Fig. 5 A disinfection fog machine and the working operational diagram

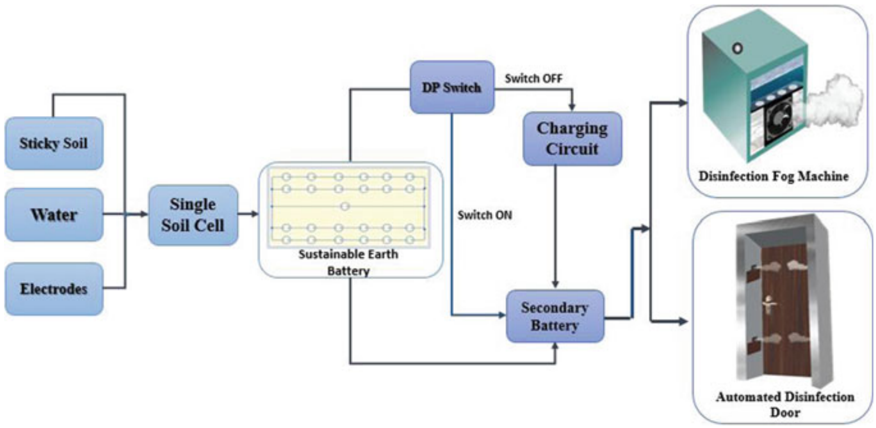


Fig. 6 The functional flowchart of the Sustainable Earth-Battery prototype

charged without external power within 2–3 hour from Sustainable Earth-Battery. A charging circuit had been implemented, which charged the secondary battery from the Earth-Battery through a D.P. Switching condition. When the D.P. switch is turned off, the Sustainable Earth-Battery charges the secondary battery. When turned on, the secondary battery connects to the O/P Load, such as a disinfector machine, to disinfect the environment. The functional flowchart of the Sustainable Earth-Battery prototype has shown in Fig. 6.

This designed prototype can run in three ways. Automatically, remotely and using voice commands, an operational flowchart as shown chronologically in Figs. 7, 8 and 9.

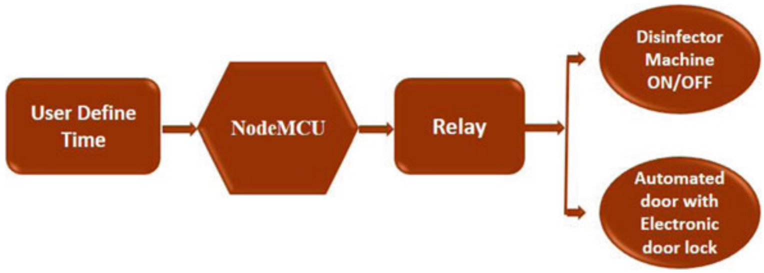


Fig. 7 Flowchart of a smart IoT-based automated sanitization system

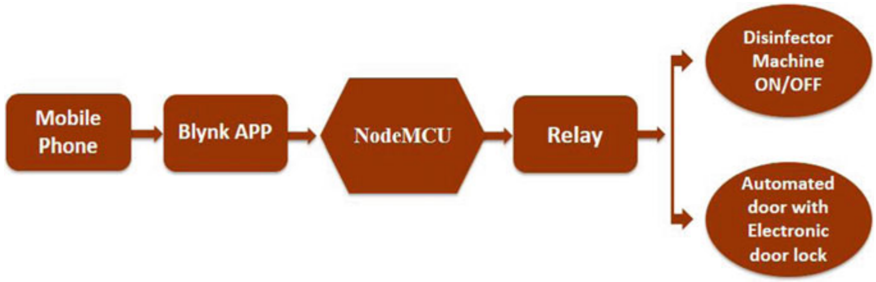


Fig. 8 Remotely controlled flowchart of the designed prototype

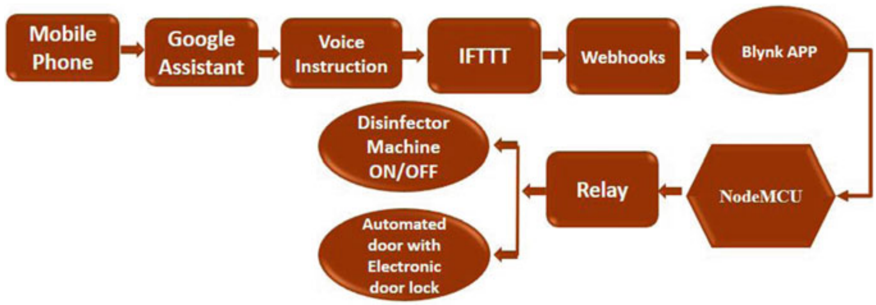


Fig. 9 Voice access controlled flowchart of the designed prototype

4.1 Designed Prototype Operational Flowcharts

- Automated Process
- Remote Access
- Voice Command Access.

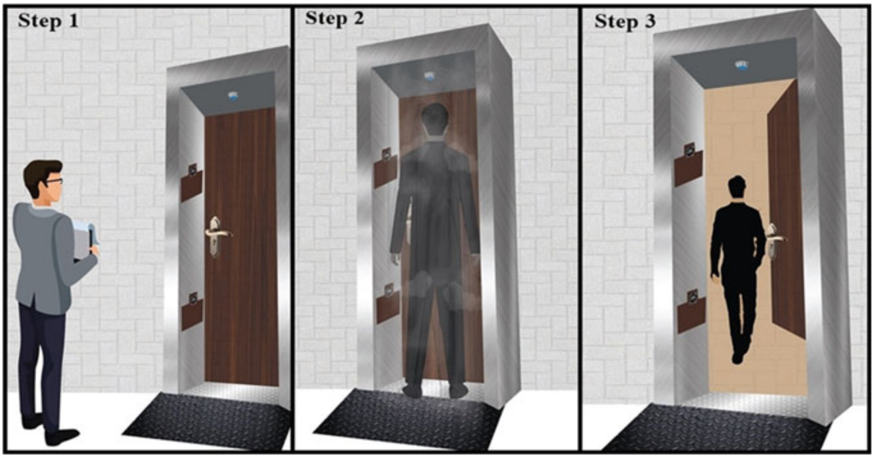


Fig. 10 Steps of the mechanism of an automated disinfector door

4.2 The Mechanism of Automated Disinfector Door

The automatic disinfector door had connected to four disinfector machines in parallel. A proximity sensor receives data whenever a person walks toward the entrance. The disinfector machine works in the meantime; the infrared temperature sensor also measures the person's body temperature. It will sanitize the person for about 10 s. After that the door will, automatically unlock if the body temperature is less than 100° Fahrenheit. There are three steps to the process, displayed in Fig. 10.

4.3 Disinfectants/Antiseptic Liquid Solution Might Use

- **Chemical**: A disinfector machine uses hydrogen disinfectant, hydrogen peroxide-based disinfectant, or a pure substance [10]. Disintegrated peroxide is the safe substitute (VPHP). Bacteria, leaves, microorganisms, and pathogens are successfully used by food grade H_2O_2 (3%) [11]. $H2O_2$ with glycerin solution is preferable as it doesn't harm human body skin.
- **Natural**: From ancient times, neem has been used in natural medicines. Locals use it as a potent disinfectant. People still use neem recipes to fight epidemics. They had used it as a tonic [12]. Boiling neem tree bark, flower leaves, and seeds could use as a disinfectant and treat skin problems. Neem, turmeric, glycerin, calendula oil, and lemon juice solution can disinfect the human body.

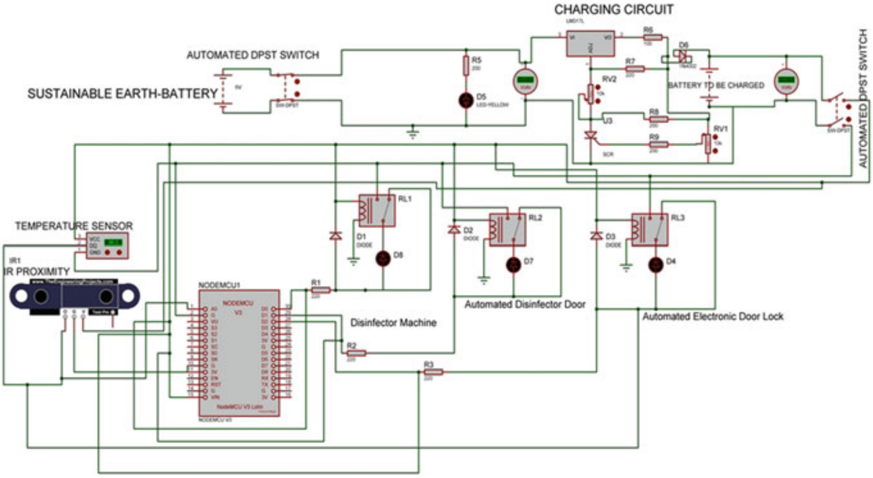


Fig. 11 Schematic diagram of the designed prototype

5 Hardware and Software-Based Implementation

5.1 Schematic Diagram

The schematic diagram of the prototype had been designed with Proteus 8 software, shown in Fig. 11. In this designed, simulated prototype, a 6 V D.C Sustainable Earth-Battery is used, the primary power source that will give power to the secondary battery, which stores charge from the source and runs this prototype continuously. The temperature sensor, I.R. proximity sensor, L.E.D.s, Resistor, DPST, voltmeter, and Battery are all connected with the NodeMCU. The necessary programming is done in Arduino IDE software.

5.2 Visualization Steps of the Hardware Implementation of IoT-Based Automated Sanitization System (Disinfector Machine)

A disinfector machine had designed to evaporate mist over a predetermined period. With this prototype, one may disinfect the surface on which it had mounted, and it is effective in the auto-sanitization cycle, killing germs, and pathogens shown in Fig. 12.

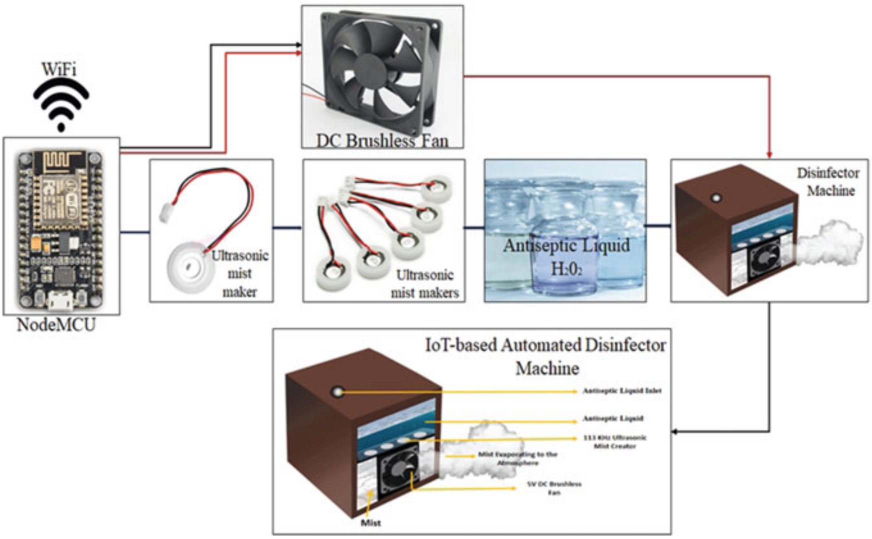


Fig. 12 Steps of the hardware implementation of IoT-based automated sanitization system

5.3 *Implementation of Automated Disinfector Door*

An automated disinfector door is a door that runs automatically when a person's movement detects by a proximity sensor and starts sanitizing process. The door has an electronic lock that automatically opens when the sanitizing process is complete and closes automatically after a few seconds. A design of an automated disinfector door has shown in Fig. 13.

5.4 *Steps of Implementation of Hardware of Sustainable Earth-Battery*

The phases of the hardware implementation of Sustainable Earth-Battery had shown in Fig. 14.

5.5 *Smart Engineering Diagram of the Designed Prototype*

Figure 15 represents the whole engineering diagram of the designed prototype.

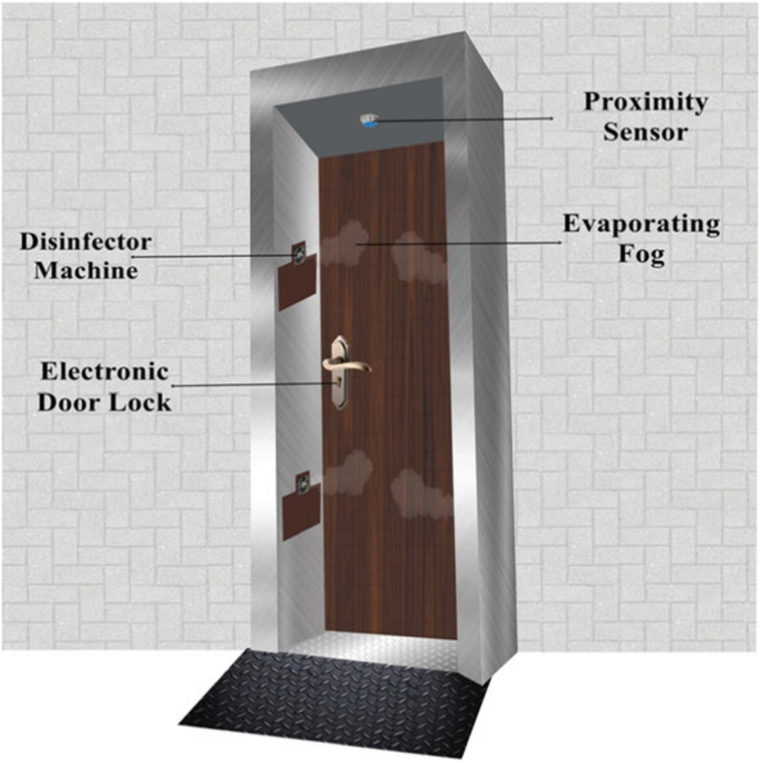


Fig. 13. 3D design of an automated disinfector door

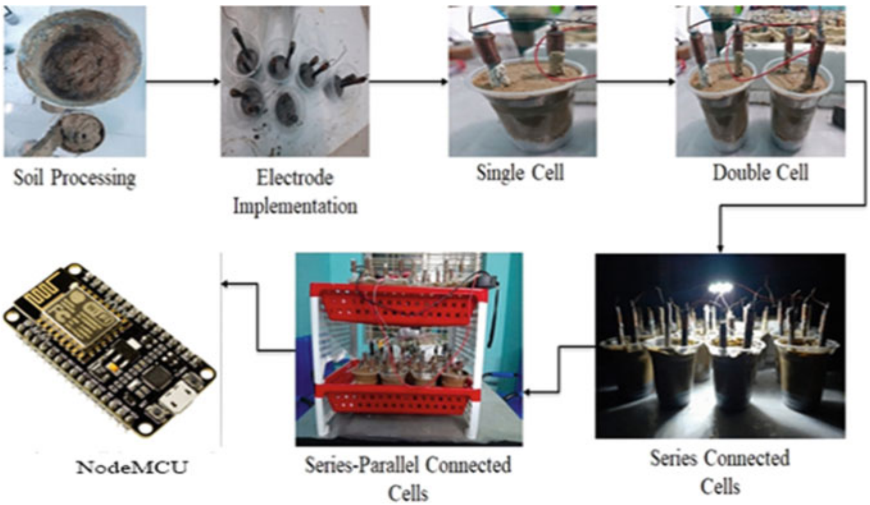


Fig. 14 Steps of Sustainable Earth-Battery implantation

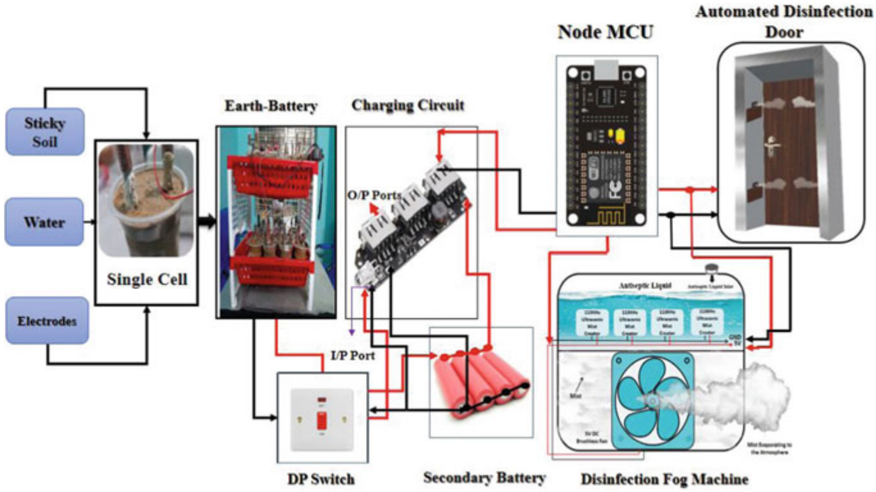


Fig. 15 Smart engineering diagram of the designed prototype

6 Result and Data Analysis

This section correctly identified the results of the prototype simulation and data used in this research. Finally, a statement on the entire prototype had described in this research. The simulation has been divided into two parts.

6.1 Charging Phase

A DPST switch transfers 6V D.C electricity from the Earth-Battery to the secondary battery, then charging started. The secondary battery is a compelling power storage device for the primary source of power—the secondary source in load mode that powers the disinfector machine, automatic door, and door lock. The 1st DPST switch, depicted in Fig.16, is off throughout this time.

6.2 On Load Phase

In 2–3 hour, Earth-Battery will charge the secondary battery, providing a constant power source for the entire system, such as a disinfector machine and automated door lock. It is a continuous process that does not require an external power source to function. The simulated design has shown in Fig.17.

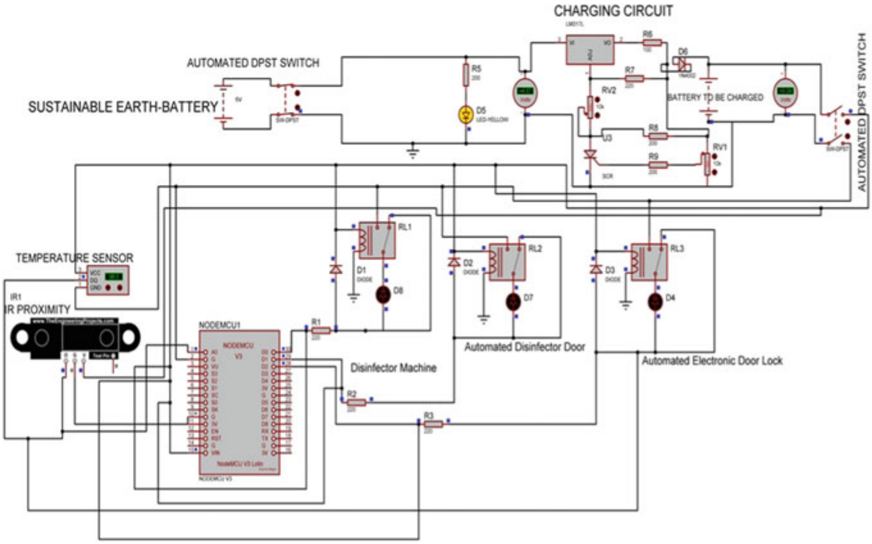


Fig. 16 Simulated design of charging phase of the designed prototype

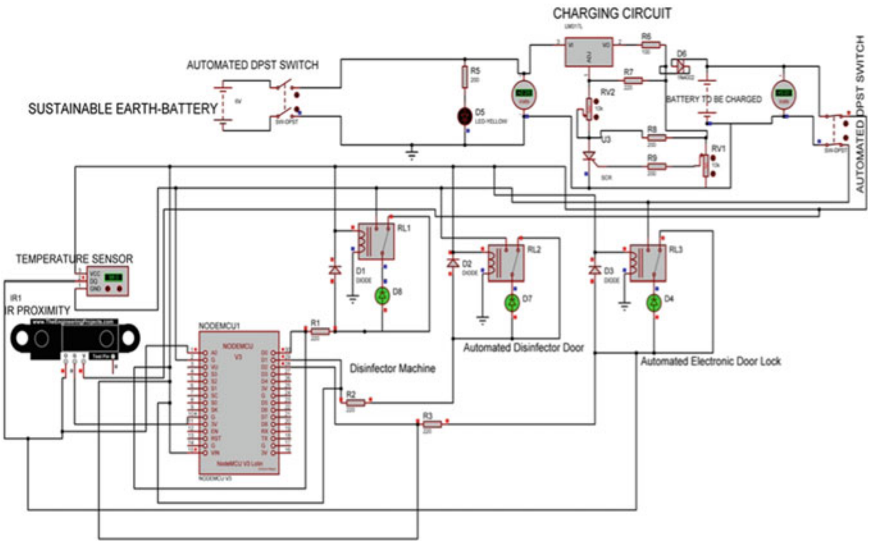


Fig. 17 On load condition of the designed, simulated prototype

6.3 Experimental Data and Findings

Each disinfector fog machine's features, shown in Table 1, suggest that it may sanitize effectively (Table 2).

Table 1 Disinfector machine analysis

No	Specifications
01	Each machine has the capability of treating 400 cft
02	The vapor fanned out over a distance of 9 ft and reached a height of roughly 7 ft
03	H_2O_2, H_2O_2 with a glycerin solution, natural disinfectant neem, and other plant herbal solutions and fungicides had used for disinfectants
04	The non-flammable and non-alcoholic disinfectants had used on the disinfector machine
05	1 L of antiseptic solution in each disinfector machine

Table 2 Operational sensor data

Sensor	Operational description
I.R. proximity sensor	It can detect human movement within a 40–50 cm range
Temperature sensor	The average temperature is 98 F, but if it exceeds 100 F, the disinfector door will not open
Disinfector door	It has four disinfector machines consisting of a 113 kHz ultrasonic Mist Maker
Sustainable Earth-Battery	6 V, 100 mA
Li-on batteries	Four 2600 maH batteries in total 10,400 mah battery charged from Sustainable Earth-Battery for continuous power supply
Charging circuit	Output: 5 V, (1–2) A, two O/P

Table 3 Comparison table analysis of Sustainable Earth-Battery prototype

Earth-Battery prototypes	Average voltage (V)	Average current (mA)	No. of cell	Voltage loss (%)
Series	24.18	25	25	5.32
Series–Parallel	6.13	100	25	4.66

Table 3 shows that Series–Parallel prototype is the perfect one.

6.4 *Naturally Auto Recharging Capability of Earth-Battery and Running Capacity of the Prototype*

See Fig. 18.

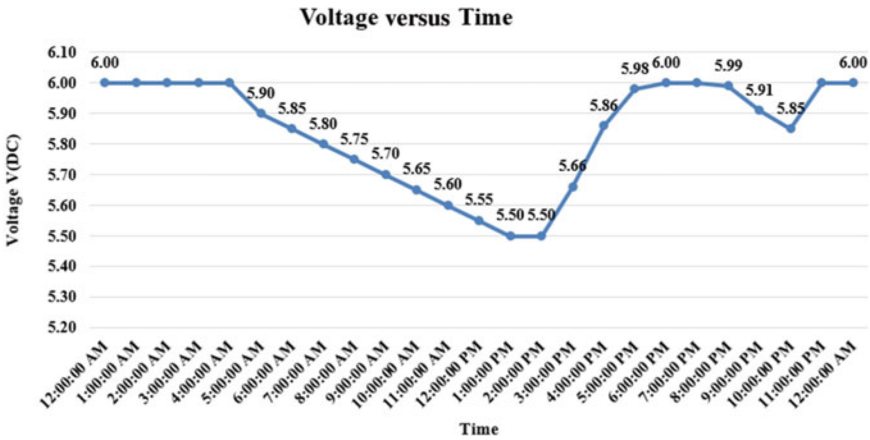


Fig. 18 Graphical representation of naturally auto recharging of the Earth-Battery

7 Future Developments and Impacts

This prototype's future applications include developing IoT technology, artificial intelligence technology, and energy waste reduction. The development of a healthier environment, the advancement of IoT and automation technologies, and the elimination of requirements and enjoyable expenses are prospective applications for this prototype. Additionally, this prototype will create new potential for future renewable and alternative energy sources based on Sustainable Earth-Battery. This prototype has a plethora of social and environmental implications. It can potentially halt the spread of the potentially fatal COVID-19 virus. It can meet the unending demand for electricity because a sustainable Earth-Battery is a cost-free source of electricity. This prototype is durable and cost-effective, and it has the potential to create new jobs and resolve the unemployment problem. It is a more modern and pleasant approach to social life. Earth-Battery energy is green energy, which does not produce damaging radiation or CO_2 into the environment. Since the prototype uses a non-flammable, non-alcoholic liquid, it will not cause a natural disaster. It is a secure technology that will increase consumer interest.

8 Applications

This study gave the latest pandemic approach. An "IoT-based Auto-Sanitization System Powered by Sustainable Earth-Battery" would help in a pandemic. This prototype might be utilized in schools, hospitals, and public transportation to promote a healthier environment while protecting home equipment. The prototype was made with cheap parts to keep costs low and increase availability. The Earth-Battery can

Fig. 19 Visual representation of possible applications in public places

be used as an alternative source. The inner surface of cars, buses, and trains can use this prototype for keeping safer environment to reduce the community transmission. This study method can reduce the epidemic. Disinfector machines are helpful everywhere to use. Schools, institutions, offices, and public venues can use it (Fig. 19). It can be utilized both indoors and outdoors. The parts of the device are inexpensive. It is possible to control via a mobile phone remotely.

9 Conclusion

Sanitization protects us against fatal viruses like Corona. Efforts have been made to halt the virus from spreading. The importance of cleanliness in this corona pandemic cannot be overstated, as the entire space was scrubbed. Using Coronavirus and advanced technologies to solve this problem is a significant hurdle. "Sustainable Earth-Battery Powered IoT-based Automatic Sanitization Device" would be a great solution and standard with renewable power sources. The Electrical, electronics, and automation sectors would profit from this prototype. This prototype is a low-cost, scalable, eco-friendly, energy-efficient self-sanitation system with health protection. In a long-term, this prototype could aid with the pandemic. It will always produce a healthy environment for everyone, epidemic or not.

Acknowledgements First and foremost, the authors wanted to express gratitude to the Almighty and their parents for their continuous support. The authors would like to thank Vice-Chancellor Dr. Carmen Z. Lamagna and the Electrical and Electronics Engineering Department of American International University-Bangladesh for their support. Thanks to the Power Cell, Power Division, Ministry of Power, Energy and Mineral Resources (MoPEMR)-Bangladesh for awarded the Sustainable Earth-Battery project, which is two times National Awarded prize winner from the Ministry and as a part of the BICCHURON-19 National Awarded Research Project of Power Cell based on Renewable Energy.

Finally, the authors want to convey gratitude to Prof. Dr. Md. Abdur Rahman and Ashraful Alam, Ali Khan Majlish and Abdur Rob Shetu from Genvio Pharma Limited for their inspiration and motivation for the research work.

References

1. Wan, Y.S.: Education during COVID-19. Brief Ideas **19**, 3–9 (2020)
2. Sayeduzzaman, M., Borno, M.I., Islam, M., Howlader, S.: A design of an IoT-based smart home with auto-sanitization system. In: Zhang, Y.D., Senjyu, T., So-In, C., Joshi, A. (eds.) Smart Trends in Computing and Communications. Lecture Notes in Networks and Systems, vol. 286. Springer, Singapore. https://doi.org/10.1007/978-981-16-4016-2_19
3. Yadav, A.K., Rajpoot, D.S., Shukla, S.S.P.: I.O.T. model-UV-based system for Sanitization of package surfaces. J. Phys. Conf. Ser. **1714**(1), 012010 (2021). I.O.P. Publishing
4. Mohammed, M.N., Arif, I.S., Al-Zubaidi, S., Bahrain, S.H.K., Sairah, A.K., Eddy, Y.: Design and development of spray disinfection system to combat coronavirus (Covid-19) using IoT based robotics technology. Revista Argentina de Clínica Psicológica **29**(5), 228 (2020)
5. Nnebedum, B., et al.: Solar-powered soil condition activated irrigation system with automated bird repellant. I.O.P. Conf. Ser. Earth Environ. Sci. **331**(1) (2019). I.O.P. Publishing
6. Belov, V.V., et al.: The Earth-Battery research as the source of renewable energy. In: International Multidisciplinary Scientific Geo Conference: SGEM 18.4.1., pp. 713–9 (2018)
7. Pakpoom, C., Bamrung, C., Padhetpoom, A.: The study of marl soil potential in electricity generation, August 2013. ISBN 978-90-78677-77-2
8. Borno, M.S.I., Sayeduzzaman, M., et al.: An empirical analysis of Sustainable Earth-Battery. Energ. Rep. **7**(Supplement 3), 144–151 (2021). ISSN 2352-4847, https://doi.org/10.1016/j.egyr.2021.06.026, https://www.sciencedirect.com/science/article/pii/S2352484721003917
9. Sayeduzzaman, M., Borno, M.S.I., Fariya, K.Y., Khan, M.T.A.: A convenient method and design for constructing an IoT-based smart automated sanitization system. In: Hybrid Intelligent Systems. HIS 2021. Lecture Notes in Networks and Systems, vol. 420. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-96305-7_11
10. A. (2018, July 15). How to Make Remote Control Door Lock at Home (D.I.Y. Electric Project). Creativity Buzz. www.creativitybuzz.org/make-remote-control-door-lock/
11. Gillis, A.S.: What is IoT (Internet of Things) and How Does it Work? IoT Agenda, TechTarget, 11 Feb. 2020. internetofthingsagenda.techtarget.com/definition/Internet-ofThings-IoT
12. Jamal, P., Mamun, A.A., Chik, W.D.W., Zainuddin, I.Z., Mirghani, M.E.S.: Development of process conditions to optimize the extraction of bio-disinfectant from Neem leaf (Azadirachta Indica). Adv. Nat. Appl. Sci. **6**(5), 682–686 (2012)

IoT-Based Decision Making of Irrigation and Fertilizer Management for Marginal Farmers in Bangladesh

Wahid Anwar, Tanvir Quader, Mudassir Zakaria, and Md. Motaharul Islam

Abstract With the current advancement of IoT and information technology, there are many solutions related to agriculture and soil condition monitoring. It is very important that these solutions need to be tailored as per the area and climate that the farmers are working in. The government of Bangladesh has also initiated many projects with specific goals and objectives to make agriculture more advanced in order to achieve the sustainable development of the country. But, there are limitations to getting efficient outcomes from these projects, as many services are not interconnected yet. This paper demonstrates a model for monitoring the soil condition of farming land, where a farmer can easily know the need for irrigation or fertilizer usage in his field. This model focuses on the affordability of the device for the farmers, and it also minimizes the utilization of devices.

Keywords IoT · Farming · Farmer · Soil · Moisture · Fertilizer

1 Introduction

IoT nowadays has become an integral part of doing everyday tasks. More and more home appliances, cars, and electronics have included IoT as an attractive feature. In this project, our goal is to provide marginal farmers with the means of technology solutions that are intelligent enough to monitor and compute optimal results for them.

W. Anwar (✉) · T. Quader · M. Zakaria · Md. Motaharul Islam
United International University, Dhaka, Bangladesh
e-mail: wanwar211097@mscse.uiu.ac.bd
URL: https://www.uiu.ac.bd/

T. Quader
e-mail: tquader211094@mscse.uiu.ac.bd

M. Zakaria
e-mail: mzakaria211096@mscse.uiu.ac.bd

Md. Motaharul Islam
e-mail: motaharul@cse.uiu.ac.bd

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_48

We say optimal results in terms of countrywide soil quality and crop wise need for irrigation and fertilizer management [3].

We plan to integrate data sources maintained by the government and from different releasing organizations. In different parts of the country, the practice of growing crops is different due to differences in climate, soil texture, and other influential factors. Based on these factors, the government of Bangladesh has divided the total arable area into 14 main regions. So we need to contextualize the national agricultural knowledge for a specific area [14].

In this paper, we expect that the Upazilla Agriculture Officer will contribute to contextualizing this knowledge for his Upazilla. Based on this Upazilla-based agriculture knowledge, we can analyze our reading and provide an intuitive answer to the farmers. This device will be handheld. Farmers will carry this to different sections of their cultivated land and take several readings. A similar technique of collecting data from several positions to recommend a suitable crop has been described in the paper [6] written by Larry Elikplim et al. When analysis is completed, the device will show the recommendations on an LCD monitor.

There are some key components which will increase the usage of this solution. Easy-to-use handheld sensor devices can help marginal farmers for monitoring crop and use of data sources from national services, like—soil and fertilizer recommendation from Soil Research Development Institute (SRDI), Fact-sheet of different crop variety from different releasing organization for example—Bangladesh Rice Research Institute (BRRI), Bangladesh Sugarcrop Research Institute (BSRI), Bangladesh Institute of Nuclear Agriculture (BINA), Agriculture Universities etc. Recommending the need of irrigation and fertilizer is widely worked on topic by the scholars as we can see in the work [13] of Alfian Ruslan et al.

We will construct our device to be a simple one, so that the farmers can easily carry it to the field. Our design does not include multiple sets of sensors, rather farmers will carry the device to different parts of their field and collect the readings. This will help reduce the cost of the overall solution. Firstly, less number of equipment will be needed. And Farmers School (10k), Farmers Club (30k), and Farmer Information and Advise Center (FIAC, 5k) can afford this easily. Secondly, farmers or entity responsible for maintaining the device will have to work on less number devices. We can see in most of the papers [9, 13, 17] the number of sensors is not considered a major issue. But, in our design we considered how to use less sensors as the amount of land our farmers own are of small sizes. This design will also increase the durability of the devices, as people will be able to easily check on its status. The major contributions to this paper are given below:

- We have proposed to integrate the data sources from Soil Research Development Institute (SRDI), fact-sheets of different crop varieties from different releasing organizations like Bangladesh Rice Research Institute (BRRI), Bangladesh Sugarcrop Research Institute (BSRI), Bangladesh Institute of Nuclear Agriculture (BINA), and agriculture universities.
- We proposed contextualizing national agricultural knowledge for a specific region.

- We have proposed to incorporate weather information for analyzing the recommendation.
- We have constructed a handheld device that will be used for soil condition sensing and showing recommendations.
- We have also proposed building an intelligent recommendation system for individual crops based on local agricultural knowledge. Once the data has been collected by the sensors, the system will analyze it to make an appropriate recommendation regarding irrigation and fertilizer usage.

The rest of the paper has been organized as follows: Sect. 2 describes the literature review and gap analysis. Section 3 demonstrates the methodology. And finally, Sect. 4 concludes the paper.

2 Literature Review

We have gone through several papers. Among them, some of them, we found, are to be very contemporary with our topic. First is the work of Larry Elikplim et al. They worked to take readings in five places and suggest suitable crops [6]. Also, they used a commercially available soil moisture sensor, FC-28. Again, in the work of Abdullah Na et al. determination of soil temperature is done using the DS18B20 sensor working on the Dallas one-wire protocol [10].

In the project "Smart agriculture to measure humidity, temperature, moisture, pH, and nutrient values of the soil using IoT" by Asadi Venkata Mutyalamma, Gopisetty Yoshitha, Althi Dakshyani, and Bachala Venkata Padmavathi, Arduino is used with various sensors to monitor the different stages of plant cropping like moisture, temperature, humidity, pH value, and nutrients of the soil [12]. They are using the Arduino Uno model with the GSM module to help with the processing, transmission, and reception of data between sensors and the microcontroller [1].

Objective of the study "Determination of soil moisture using various sensors for irrigation water management" by Praveen Barapatre and Jayantilal N. Pate is to examine and analyze prototyping an organized calibration and working of various soil moisture sensors. Outputs of this study indicate IoT-based soil moisture detection is an effective method that provides reliable front data that can be used to adapt and improve irrigation and precision farming methods [2, 4, 5].

In the journal "2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017" by Manish Bhimrao Giri and Ravi Singh Pippal, they showed that wireless sensor networks provide a very optimal solution for water distribution using interpolation methods [7].

The study, "The IoT-Based Monitoring Systems for Humidity and Soil Acidity Using Wireless Communication", aims to create a pH and humidity monitoring system for agriculture's soil with wireless sensor network technology based on IoT. The monitoring system is able to display pH and soil moisture values in real time. Compared to commercial soil analyzers, the average error value of the soil pH sensor is equal to 1.66% and the YL69 sensor error average is 1% [8].

In another work, we evaluated "IoT-Based Soil Condition Monitoring Framework" by Selvakumar Manickam, discussed how IoT can help determine suitable crop for a land and also managing nutrient level by applying appropriate fertilizer [9]. Various sensors were used to measure temperature, moisture and light, humidity, and pH value for this project. The MCP3204 ADC was used to collect the data from sensors and then sent to cloud.

In the paper "Development and application of cell phone-based Internet of things (IoT) systems for soil moisture monitoring", Jose O. Payero, Michael W. Marshall, Bhupinder S. Farmaha, Rebecca Hitchcock Davis, and Ali Mirzakhani Nafchi developed and field tested affordable cell phone-based IoT systems for monitoring soil moisture. These IoT systems would be accurate, affordable for small farmers, robust under normal field conditions, reliable, and easy to use [11].

In the project "IoT soil monitoring based on lora module for oil palm plantation", Ahmad Alfian Ruslan, Shafina Mohamed Salleh, Sharifah Fatmadiana, Wan Muhamad Hatta, Aznida Abu, and Bakar Sajak worked to provide a solution to overcome the manual monitoring system used by the workers for a very long time. Their system will ensure the increased productivity and efficiency of the planter at the tip of their fingers [13].

Mrs. Shwetha designed and developed the system, which measures water quality using sensors such as pH sensors, temperature sensors, moisture sensors, and turbidity sensors in conjunction with the Raspberry Pi. The measured quality by sensors is transmitted to the Raspberry Pi, and then it is sent to the controlling center through the Internet. This IoT system allows the user to access the data from the database through the website and provides the advantages of improved accuracy, efficiency, and low cost [15].

In the work of P. Divya Vani and K. Raghavendra Rao, they showed how a low-cost soil moisture sensor with a CC3200 LaunchPad can be used to measure the soil condition [16]. They collected two types of soil, namely red and black, from the field and carried out experiments at room temperature. For moisture level measurement, the FC-28 soil moisture sensor was used. Finally, the data was uploaded to AT & Ts' M2X Cloud. This data is made available through web and mobile applications.

In the project "IoT-based soil moisture measuring system for Indian agriculture", V. Vanitha, Rajat Kumar Dwibedi, Ben Painadthu, Kiran Venugopal, and Aby Varghese Thomas provided a soil moisture sensor system to test and monitor soil humidity with the Node MCU system. The two major parts are the moisture measuring system and the mobile application for better user experience. The data is accessed by the mobile application whenever required. The mobile application would be made generic for the other sensors, and the required security would be provided at the user level. They also created a cloud-based platform that allows users to query soil moisture of any land resource from their smartphones [17].

Most of the papers we have gone through have mainly focused on the IoT solutions for irrigation or gathering nutrients needed in a field. But, it is very important to base the irrigation suggestion on the soil texture of our country. Also, the need for nutrients must be based on the variety of crops our farmers produce. In Fig. 1, we have shown how we plan to fill the gap by using IoT with national data sources.

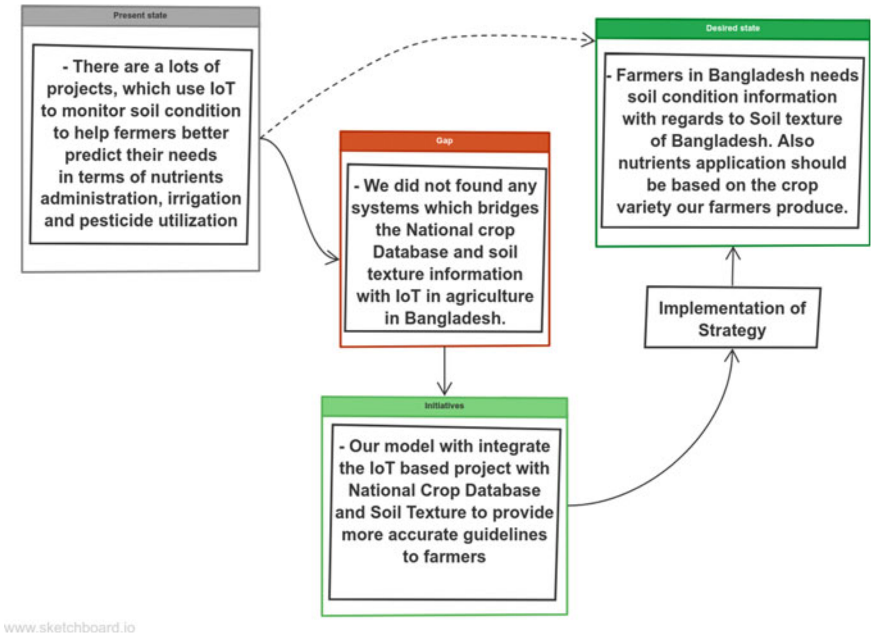


Fig. 1 This figure describes gap analysis

3 Methodology

In this project, we will build a simple handheld IoT device that can read soil conditions and send these readings to a server for analysis. When readings are received by the server, it will initiate a process for recommending farmers. This process will incorporate data from the Soil Research Development Institute (SRDI) and factsheets of different crop varieties from different releasing organizations, for example, Bangladesh Rice Research Institute (BRRI), Bangladesh Sugarcrop Research Institute (BSRI), Bangladesh Institute of Nuclear Agriculture (BINA), and agriculture universities. Based on these national data sets formulated for specific crops, soil texture, or areas, we will be able to provide a precise recommendation for the farmer.

In this process, the Upazilla Agriculture Officer will play a vital role. These officials serve as the primary points of contact between farmers and the scientific community. They know the area's specific crop cultivation practices better than anyone else. If the nation's wide agricultural knowledge is not used with the adjustment needed for local context, we will fail to reap the full potential. Agricultural universities across the country are the first to disseminate the results of their scientific research to farmers. They conduct workshops or demonstrations in the field to make the farmers aware of this information. Unless the agricultural officers communicate the knowledge of the scholarly community to the farmers, it will have little value. So, for accurate soil and crop information, these national databases are intriguingly dependent on the refinement of agriculture offices.

In the current scenario, farmers have to go through 8–10 steps to know the need for fertilizer on their land. They collect samples and send them to the agricultural office to analyze. After getting soil condition results, agriculture extension officers check the crop fact-sheets and fertilizer recommendation database to evaluate the results.

In Fig. 2, we can see the steps farmers have to follow to get the recommendations. Finally, the farmers are provided with a recommendation based on the result (Fig. 3).

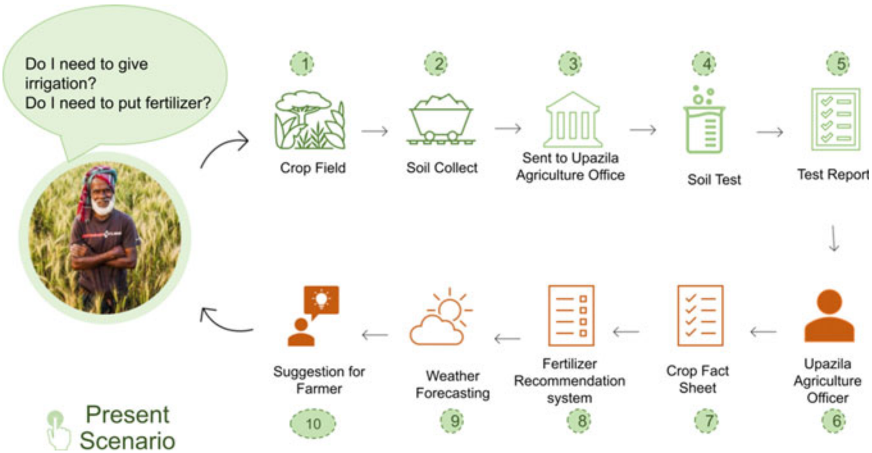


Fig. 2 Current scenario how farmers now avail this service

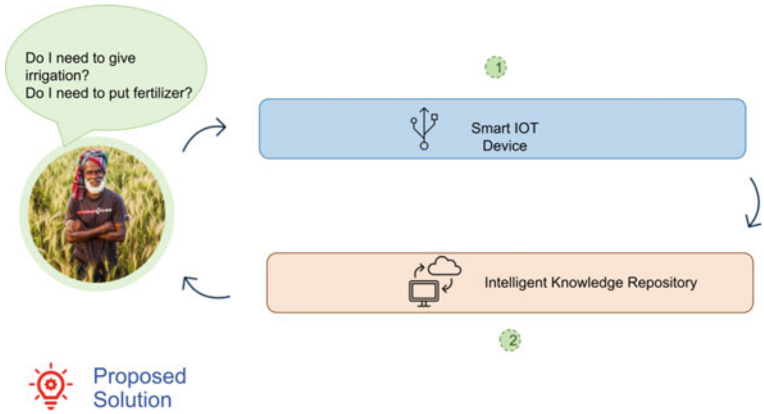


Fig. 3 Proposed solution to replace current process

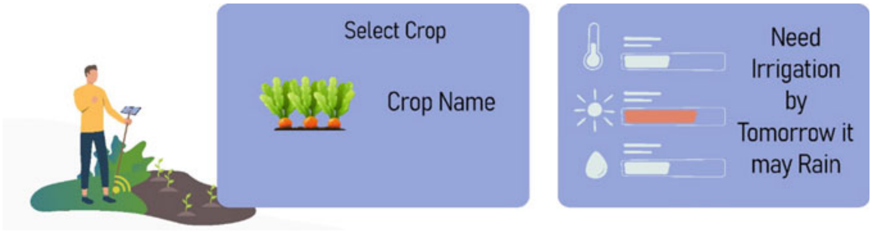


Fig. 4 Sample process for moisture detection output

In Fig. 4, we show a sample process of the outcome of the service we are proposing to replace the current one. Here, we have shown how the need for irrigation for the field may be communicated with the farmers. Similarly, the need for specific fertilizer applications may also be shown in the display.

The handheld device will consist of two sensors and a monitor. There will be a soil NPK sensor in the device which will provide the soil nutrition requirement. Another soil humidity and soil moisture detection sensor will be used to detect if the field needs irrigation or not. There will be a simple monitor to display the recommendations provided by the server. In Fig. 5, we have demonstrated the process flow of how the device will perform.

In Fig. 6, it is shown how the analysis of suggestions will take place with respect to the sensors' readings. The sensor reading data will be sent to the server after being collected from the field using the microcontroller's WiFi module. Then the server will initiate the processing of formulating the suggestions. When ready, this data will be sent back to the device from the server. Finally, the device will display the suggestions to the farmers. The suggestions made from server side analysis will consider the soil texture of the particular area, crop variety specific information, and weather information. It is vital to consider this information for the solution, which will provide an edge to the farmers while making decisions regarding irrigation and fertilizer administration.

The NPK sensor will provide the nitrogen, phosphorous, and potassium readings of the soil. It also helps to measure the fertility of the soil and facilitate the systematic assessment of its condition. It can be buried in the soil for a long time. The NPK sensor has a high-quality probe, electrolytic resistance, rust resistance, salt, and alkali corrosion resistance, to ensure the long-term operation of the probe part. So, this sensor is suitable for the detection of all kinds of soil, like alkaline soil, substrate soil, acid soil, seedling bed soil, and coconut bran soil.

It does not require any chemical reagent and can be used with any microcontroller as it has a fast response speed, high measurement accuracy, and good interchangeability. Since it has a Modbus communication port, we cannot use the sensor directly with the microcontroller. Hence, we need a Modbus Module like "RS485/MAX485" to connect the sensor to the microcontroller. The NPK sensor operates from 9 to 24 V and has very low power consumption. It is accurate to within 2%. The nitrogen, phosphorous, and potassium measuring resolution is up to 1 mg/kg (mg/l). We can

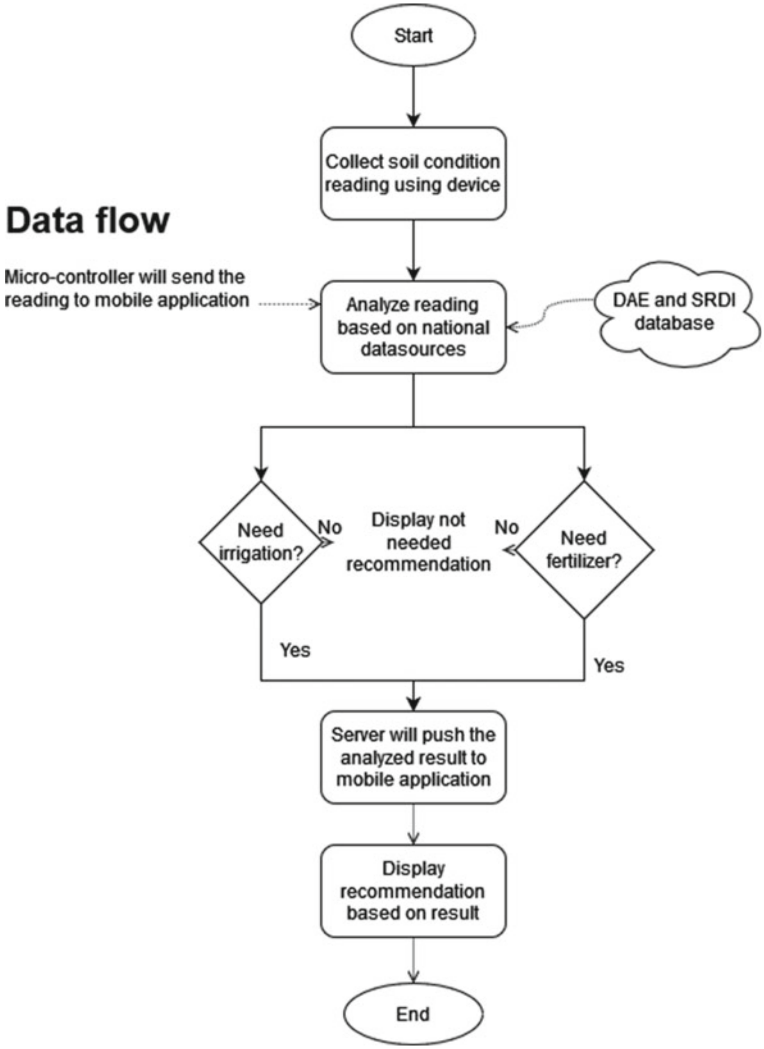


Fig. 5 This figure describes the flow of the process

create our own Arduino soil NPK meter or any cloud IoT-based soil nutrient content monitoring system using the soil NPK sensor. In Fig. 7, we have shown a sample NPK sensor we will use. If the initial project is a success, we will add more soil nutrient sensors.

As we mentioned earlier, we will provide irrigation recommendations to farmers. For that, we will use the soil humidity sensor along with the previous one. A humidity sensor is shown in Fig. 7. Our device will include a simple monitor to display the recommendation to the farmers. A sample monitor is shown in Fig. 7.

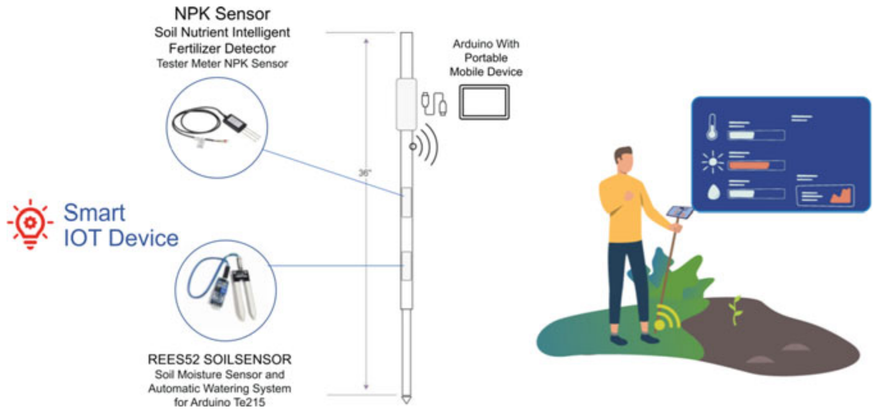


Fig. 6 Different components of the device

Fig. 7 yl-69 soil humidity moisture sensor, soil NPK sensor, and LCD display of the device

We will integrate our readings from the IoT device with the Bangladesh Rice Knowledge Bank for specific needs for the varieties our farmers produce. Another integration with the national fertilizer administration info database will be done. Based on the readings from our device, accurate fertilizer administration guidelines will be formulated from this data source.

We shall revisit the proposition of using a mobile application along with the proposed one in the future.

4 Conclusion

IoT-based agricultural solutions are making farmers more able to meet the requirements of this modern era. The resources used in agriculture will not change much, but the techniques used will evolve over time. More mechanical usage may help get a better yield, but information technology is playing a vital role in making the process more efficient.

Our proposed model is an initial step where we try to bring the power of IoT and intelligence from data sources like the National Crop Database and Fertilizer Administration Database. This model can be further enhanced by adopting more sensors and intelligence sources if necessary.

References

1. Smart agriculture to measure humidity, temperature, moisture, ph. and nutrient values of the soil using IoT. Int. J. Eng. Adv. Technol. **9**, 394–398 (6 2020)
2. Ahmed, S., Hossain, M., Kaiser, M.S., Noor, M.B.T., Mahmud, M., Chakraborty, C., et al.: Artificial intelligence and machine learning for ensuring security in smart cities. In: Data-Driven Mining, Learning and Analytics for Secured Smart Cities, pp. 23–47. Springer (2021)
3. Al-Amin, S., Sharkar, S.R., Kaiser, M.S., Biswas, M.: Towards a blockchain-based supply chain management for e-agro business system. In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering, pp. 329–339. Springer (2021)
4. Barapatre, P., Patel, J.N.: Determination of soil moisture using various sensors forirrigation water management
5. Chowdhury, F.H., Raisa, R.A., Azad, M., Uddin, S., Kaiser, M.S., Mahmud, M.: Low-cost stand-alone smart irrigation system: a case study. In: Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering, pp. 349–356. Springer (2022)
6. Elikplim, L., Akpalu, K., Owusuaa, R.M., Gyening, M., Yakubu, O., Jr, I.N.: A novel soil moisture, temperature and humidity measuring system—an IoT approach. J. Theor. Appl. Inform. Technol. **15** (2021)
7. Giri, M.B., Pippal, R.S.: Use of linear interpolation for automated drip irrigation system in agriculture using wireless sensor network, pp. 1599–1603. Institute of Electrical and Electronics Engineers Inc. (6 2018)
8. Kamelia, L., Nugraha, Y.S., Effendi, M.R., Priatna, T.: The IoT-based monitoring systems for humidity and soil acidity using wireless communication. In: 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT), pp. 1–4 (2019)
9. Manickam, S.: IoT-based soil condition monitoring framework
10. Na, A., Isaac, W., Varshney, S., Khan, E.: An IoT based system for remote monitoring of soil characteristics, pp. 316–320. Institute of Electrical and Electronics Engineers Inc. (2 2017)
11. Payero, J.O., Marshall, M.W., Farmaha, B.S., Davis, R.H., Nafchi, A.M.: Development and application of cell-phone-based internet of things (IoT) systems for soil moisture monitoring. Agric. Sci. **12**, 549–564 (2021)
12. Rahman, A., Roy, S., Kaiser, M.S., Islam, M.S.: A lightweight multi-tier S-MQTT framework to secure communication between low-end IoT nodes. In: 2018 5th International Conference on Networking, Systems and Security (NSysS), pp. 1–6. IEEE (2018)
13. Ruslan, A.A., Salleh, S.M., Fatmadiana, S., Hatta, W.M., Abu, A., Sajak, B.: IoT soil monitoring based on loRa module for oil palm plantation
14. Sadia, S., Propa, M.B., Al Mamun, K.S., Kaiser, M.S.: A fruit cultivation recommendation system based on Pearson's correlation co-efficient. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), pp. 361–365. IEEE (2021)
15. Shwetha, M.: Design and development of IoT device to measure the quality of water and water content in soil
16. Vani, P.D., Rao, K.R.: Measurement and monitoring of soil moisture using cloud IoT and android system. Indian J. Sci. Technol. **9** (2016)
17. Vanitha, V., Dwibedi, R.K., Painadthu, B., Venugopal, K., Thomas, A.V.: IoT Based Soil Moisture Measuring System for Indian Agriculture, vol. 1964. IOP Publishing Ltd (2021)

Monitoring and Controlling of Smart Irrigation System

Md. Abu Sayem, Md. Siddiqur Rahman, and Innika Khatun

Abstract The goal of this project is to create a complete agricultural system that will enable our farmers to use the least expensive technologies. Sensors are utilized for the perception of the environmental conditions encompassing the crop whose outputs are obtained on Associate in nursing humanoid primarily based mobile application likewise as uploaded on the cloud. A record of this information will be maintained that may well be used for the long-run reference, i.e., within the next cropping season, enhancing the development of crop production. This paper shows the prudent utilization of the Internet of Things for typical farming. It demonstrates the utilization of NodeMCU ESP8266-based observed and controlled shrewd water system that is both cost effective and simple. Farmers may irrigate their fields more easily with the help of an automatic irrigation system. A cation concentration detector, water flow detector, temperature locator, and soil condition identifier all exist independently in this savvy water system, and these sensors microcontroller drives the pump motor. Remotely on the Internet through the ESP8266 Wi-Fi unit, the information was obtained by NodeMCU, which then passed. These are scanned using sent data and board IoT. This empowers the far-off instrument through a safe web association with the client or user. The actual time values and reference values of several parameters required by crops have been generated for an Internet website. Users can use the Internet service to control water pumps as well as keep an eye on preset reference values that might help farmers enhance productivity with higher-quality crops.

Keywords Smart irrigation · NodeMCU ESP8266 · Soil moisture sensor · Wi-Fi · Mobile or desktop screen · Web server

Md. Abu Sayem (✉) · Md. Siddiqur Rahman
Daffodil International University, Dhaka, Bangladesh
e-mail: abu15-10146@diu.edu.bd

Md. Siddiqur Rahman
e-mail: siddique33-3802@diu.edu.bd

I. Khatun
International University of Business Agriculture and Technology, Dhaka, Bangladesh
e-mail: 15303091@iubat.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_49

1 Introduction

In numerous nations, the principle type of revenue relies upon rural advancement. As the population increases, increased agricultural development is needed. For ideal plants, an appropriate amount of water is required, and at the same time, water is vital for a nutritious life. According to agricultural land, approximately 40% irrigation and 85% fresh water are used in Bangladesh. So water should be used wisely because watering in areas consumes the largest amount of water. The traditional way manual irrigation system is inept. So excess water is a greater possibility. To overcome this problem, you need to use an advanced irrigation method. It has been stated that the end of plant plants can be 30,000, deteriorate, and lead to diseases of fungal soil [1]. That's why we have to invent an agricultural system for irrigation. Here, we use automated Internet of Things (IoT) technology to solve any problem in any field. The Internet of Things simply means taking everything in the world and communicating with each other over the Internet. It is a network of devices that connect a man to a man, a man to a machine, and a machine to a machine [2].

We powered the entire system with solar energy and programmed it in such a way that the engine operated automatically. The innovation process of agricultural irrigation connects the soil moisture sensor to the Wi-Fi system. The system then sends data to the mobile application via programming, where we can get information anywhere. Programming is done so that the engine is on and off, and once upon a time, we could also get information from the engine. The Wi-Fi module converts the analog signal into a digital signal and then directs it to the engine and mobile application. If you want the constant reading of the soil moisture sensor, then we just connect the analog pin sensor. This pin is analog to a digital adapter (ADC) pin. In another case, we attach the sensor to the unit with a digital pin, where the upper and lower threshold limit is agreed upon. Again through programming and hardware components, the digital signal is sent to the engine pin. Now, if the sensor says that the moisturizing edge of the soil is equal to the minimum threshold or below the minimum threshold, the engine turns on, and in another case, the engine shuts down if the level of moisture in the soil is equal to the upper threshold limit [3]. Our primary goal is to maintain the level of moisturizing between the minimum and upper limits.

2 Literature Reviews

Solar-Powered Automated Irrigation System

In this paper, the authors focused on the floating rate motorized microcontroller-based irrigation system concept as the main emphasis work. The only power source utilized to operate the entire system is solar power. On the paddy field, sensors are positioned. These sensors continuously sense the water level and provide the farmer with a message with the water level information. Farmers can learn about the water level without going to the paddy fields. A farmer can operate the motor from afar by sending a message from his cell phone based on the water level [4].

Automatic System of Plant Irrigation

By utilizing an automated irrigation system designed to improve water consumption for agricultural crops, the goal of this article is to lessen the farmer's manual involvement. The automatic irrigation system that measures the soil's moisture content and turns the pump when the electricity is on is the subject of this study. In this project, an L293D module and a soil moisture sensor are used to automate the control of farm irrigation and soil moisture. When the electricity is on, this automatic irrigation system detects the soil's moisture content and automatically turns the pump on [5].

Arduino-Based Smart Irrigation System

This system defines the right distribution of pesticides on agricultural land to control the disease. This device incorporates a sensor affixed to the plant as well as an analysis instrument. Temperature sensors, humidity sensors, motion sensors, light sensors, vibration sensors, and total UV sensors are examples of many sensors. A sensor is a type of analytical tool that examines patterns for the presence of a specific substance. Humidity and temperature value reaction detection by multiple sensors taken from Arduino exceeds GSM usage on mobile farmers [6] (Fig. 1).

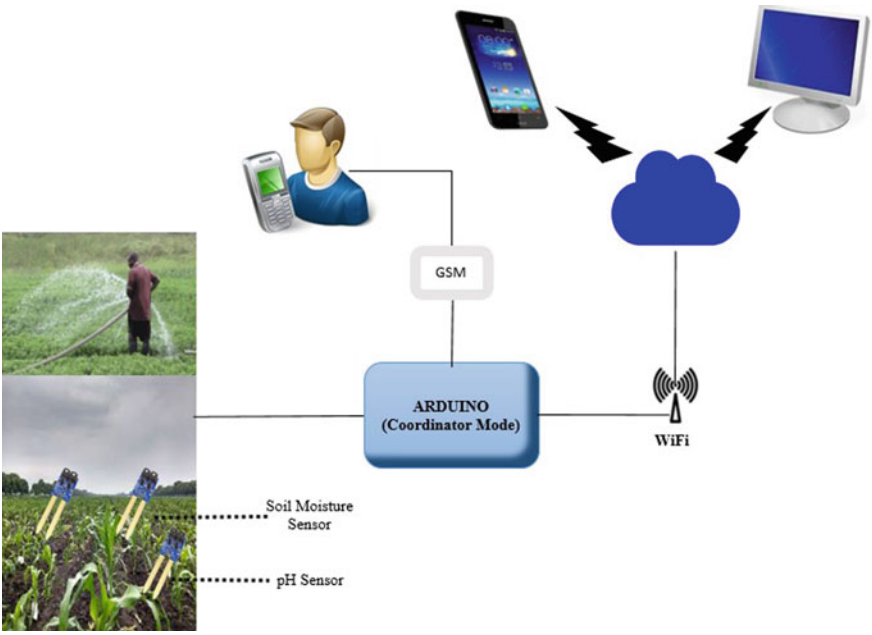


Fig. 1 Arduino-based smart irrigation system [6]

Automatic Irrigation System on Sensing Soil Moisture Content

The research also intends to create a mechanism that turns the engine on and off automatically based on humidity. Given this, enough information about the water source cannot be found, and the approach to controlling water withdrawal from supplies sufficient information was not seen about the energy supply used in this research [7].

Management System of Automatic Irrigation with Sensors

The main focus of this research is to develop a technology that can help regulate water by sensing humidity. Sensors placed on the ground throughout the globe will provide information on the need for water, which will be organized similarly. Simultaneously, a mechanical approach to the water of the tanker that would be filled when it was empty was set up [8].

Automatic Irrigation Methods for Detecting Soil Humidity

The authors of this research demonstrate how automated irrigation methods can be utilized to apply automated irrigation that is appropriate for agriculture and eliminate operational errors brought on by staff members. So the Arduino Council (ATmega328 tiny controller), which is an input marker aggregation of the shifting humidity conditions through a unique moisture detecting mechanism, is responsible for transferring the project. In order to protect valuable soil and reduce the quantity of water needed to irrigate crops, the project highlights the use of moisture sensors to control and adjust the amount of water used in crops [9].

3 Analysis of the System Components

The hardware design of the smart irrigation system consists of two parts. The electronic section is the first one and the body section is another one. The electronic section consists of several well-designed circuits containing some of the electronics. The hardware section consists of the smart irrigation system.

The system proposed is composed of the following components: NodeMCU ESP8266, Adapter 12v, Pump motor (Small Size), Relay 5v, Transistor BC547, Diode 1N4007, Jumper Wire, Pipe for the Pump motor, On/Off Switch, Capacitive Soil Moisture Sensor, L298N motor driver, and Breadboard.

4 Hardware Development and System Design

We discussed the development system and design system. We will cover the block diagram, proposed system, methodology, system description, and also working system (Fig. 2).

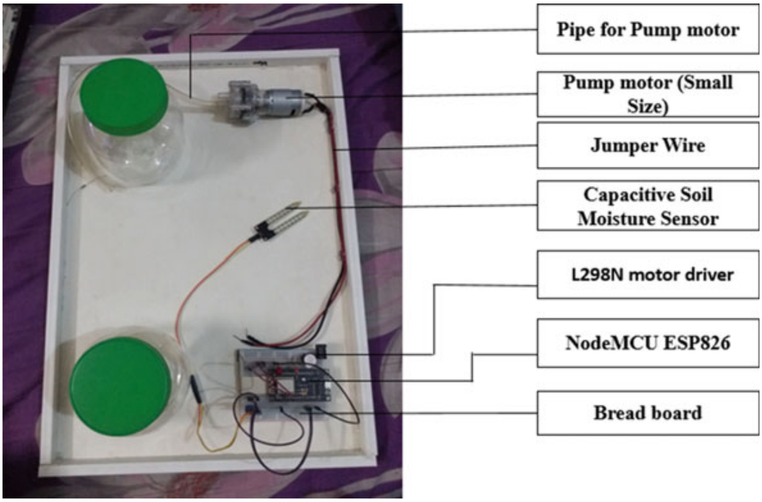


Fig. 2 Hardware design

In the IoT-based undertaking, we find out with regards to brilliant cultivating and mechanized water system frameworks with the NodeMCU ESP8266. Horticulture has a basic part in the improvement of rural countries. The government's policy of "social and economic development" is to provide a comprehensive and comprehensive system of social and economic development. Smart agriculture is the only way to solve this problem by updating current traditional agricultural processes. As a result, agriculture is smartly automating and utilizing IoT technologies. IoT empowers different applications to screen crop development and choice, support programmed water system goals, and so on. Crop productivity must be modernized and improved; ESP8266 system based on monitoring and controlling of smart irrigation is proposed.

4.1 Proposed System

The results of humidity, temperature, and threshold level can be calculated through sensors used in the project. Soil parameters can be analyzed, and nutrients for the soil can be calculated. The need for water supply from the soil can be calculated, and therefore, proper irrigation is done through smart techniques.

The artificial way to provide water for crops in the fields is irrigation. A water shortage has been advised in the current environment owing to expanded double-dealing to foster another innovation that might save water from being squandered, and since farming is the most practical business, an irrigation system would be a smart approach to reduce water loss (Fig. 3).

Water distribution is becoming increasingly constrained, and water management has become increasingly vital for irrigated agriculture's long-term viability. The

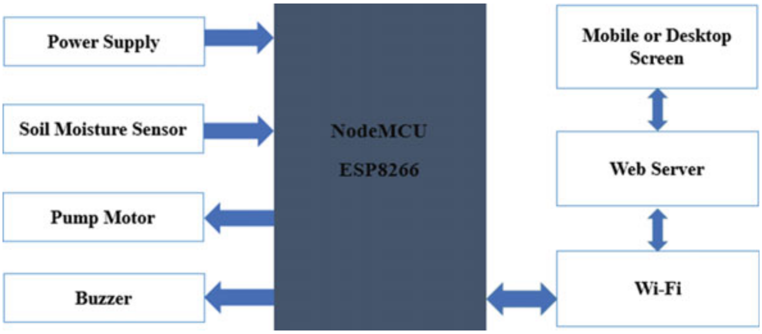


Fig. 3 Block diagram of monitoring and controlling of the smart irrigation system

objective is to make the water system framework savvy, independent, and proficient to enhance crop water conveyance while lessening manual mediation. It monitors soil, climate, and drought conditions, as well as autonomous plant water consumption and water program adaption. As a result, the shrewd water system has turned into a significant concern for the framework, with smart devices being supplied to farmers to assist them in producing good harvests. Bangladesh is mostly an agricultural country with abundant water resources. However, population growth and overuse have resulted in a situation where water demand and availability are at an all-time high. Several sensors are employed in smart irrigation systems to provide land updates to farmers. The water required for the region is calculated using soil moisture, water flow, and temperature. Because crops are temperature sensitive, the dirt dampness sensor decides the amount of dampness in the field to forestall desertification of water. Temperature sensors monitor crop warmth.

4.2 Circuit Design

The 3.3 V power supply terminal of ESP8266 is used to operate the soil moisture sensor. The VCC and ground pin of the sensor are connected to the VCC and ground pin of the microcontroller. Also, the L298N motor driver is connected to the microcontroller (Fig. 4).

4.3 System Description and Algorithm

NodeMCU is the scholarly piece of the framework. It is an open-source electronic stage where complex contraptions models with equipment and programming are planned, created, and tried. NodeMCU collaborates with the client through a site and makes a server that can get and move information to the site from the Wi-Fi ESP8266

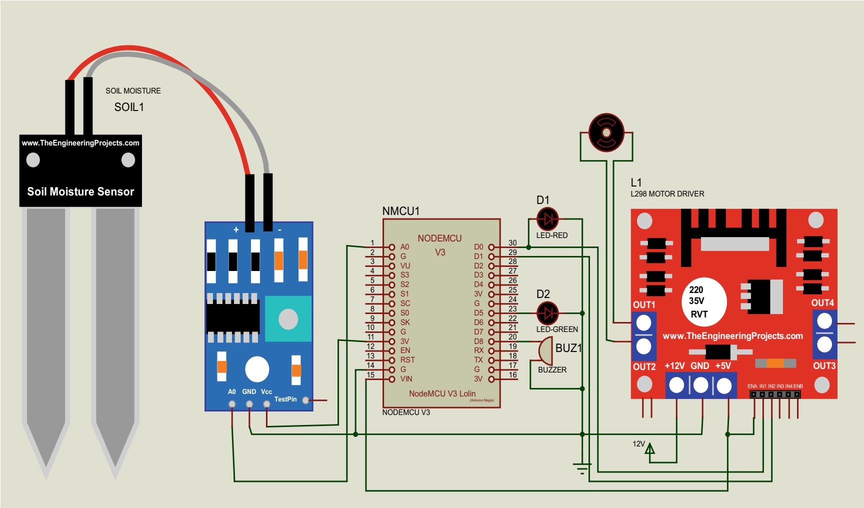


Fig. 4 Circuit design of monitoring and controlling of the smart irrigation system

(NodeMCU microcontroller association). NodeMCU—ESP8266—Server—Client or User.

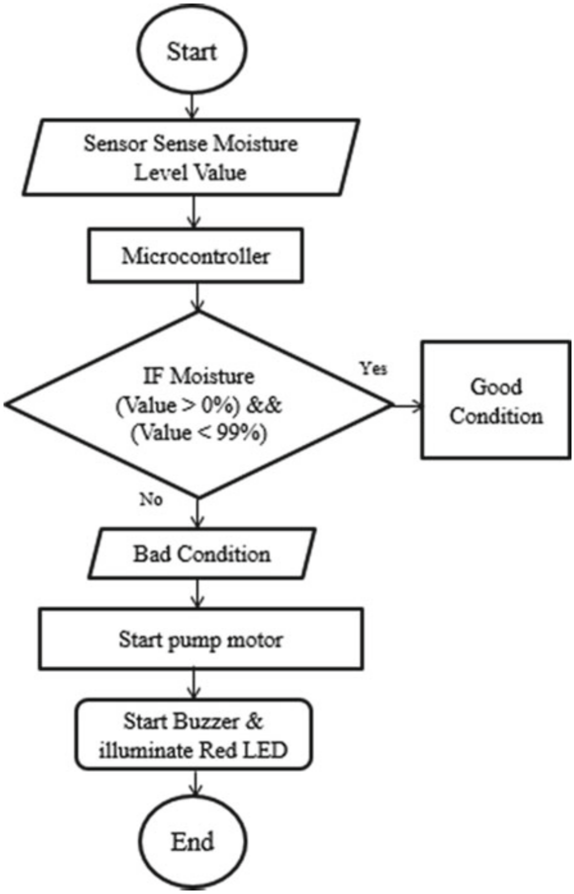
To obtain the desired output, the accurate rule is required. The algorithm is made up of the following steps:

- Step 1: Power supply turn on for the NodeMCU, sensors, and other devices.
- Step 2: Turn on the system, which contains sensors, Wi-Fi modules ESP8266, and a user interface.
- Step 3: From the configuration file, translate the system instructions and run it according to the files.
- Step 4: If the user needs ongoing alerting or monitoring, then read the sensor's data and analyze it to see.
- Step 5: Use 8266 ESP to send data to the user and wait for a response.
- Step 6: If yes, examine the irrigation pumps' case switch and record the current status in the system configuration file. The user can choose an irrigation pump by transferring particular orders.

4.4 Flowchart

See Fig. 5.

Fig. 5 Flowchart of monitoring and controlling of the smart irrigation system



4.5 Methodology

The system method is controlled through the implementation of the prototype machine that works automatically and through the mobile application. Drawing a timeline and reading relevant functions will be the first stage in the prototype format. After weighing the advantages and downsides of prior research on automated irrigation systems, we can start the implementation plan and method of automation.

Finally, we discussed the development system and design system. We discussed the block diagram, the suggested system, the methodology, the algorithm and system description, as well as the actual working system. The project showcases the design of IoT-based smart irrigation system. The proposed system could limit farmers' efforts and provide higher fields. It additionally conserves water for irrigation by police work sensors within the right position higher than the soil level. This work has shown that plants are still able to maintain low humidity levels when the temperature is moderate. Analysis of more than one parameter created this system efficiency for field management.

5 Results and Discussions

We interface all devices as well as software via wire and through the Internet and NodeMCU-ESP8266 connection. We get notifications on mobile and automatically activated them according to pump requirements. As such, if there is a digital connection, there is moisture in the soil under or equal to the minimum, then the pump motor turns on automatically and if it reaches the maximum top, the pump stops automatically. Both conditions are shown in Figs. 6 and 7 (Fig. 8).

The project is ready to use. As a result of its ease of use, maintenance, and low cost, the ESP8266 wireless communication technology was adopted for use on the web. Water demand and reliability will be accurately monitored and controlled by this system. The user can display about the information of sensors from any place on the planet in nanoseconds, which is gainful to the client. Besides, this plan utilizes the ESP8266 Wi-Fi module and the NodeMCU microcontroller, which diminishes energy utilization by expanding the system's life across wide regions with low introductory speculation.

Fig. 6 IP address connection (Device-I)

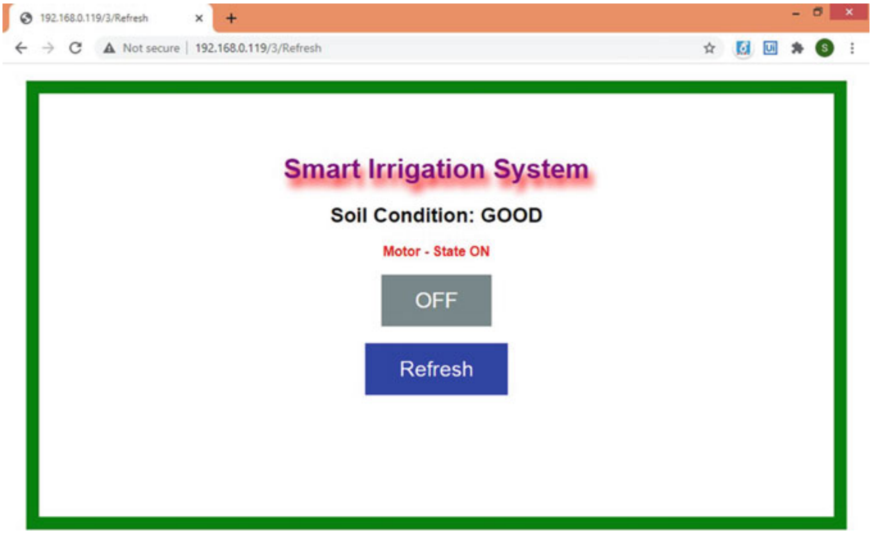


Fig. 7 IP address connection (Device-II)

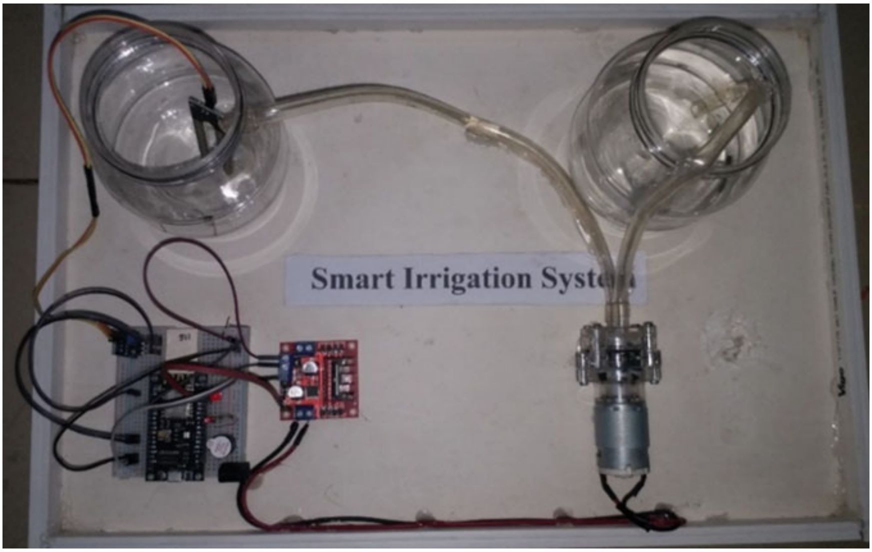


Fig. 8 Final project (top view)

6 Conclusions

In this IoT-based project, we learned about smart farming and smart irrigation systems with NodeMCU ESP8266. Agriculture has a critical part in the process of agricultural nations. Agricultural issues had always hampered the country's development. Smart agriculture is the only way to solve this problem by updating current traditional agricultural processes; as a result, agriculture is smartly automating and utilizing IoT technologies. Internet of Things (IoT) allows for a variety of applications to monitor crop growth and selection, support smart irrigation resolution, etc. To modernize and improve crop productivity, an ESP8266 system for IoT-based monitoring and controlling of smart irrigation is proposed.

6.1 Limitations of the Work

The monitoring and controlling of the smart irrigation system is a bit expensive. Depending on the size of your property, you will need more systems. Of course, savings on water bills will result in lower costs. If you want to use this system to water the garden, it is better to fix it underground before planting. The smart irrigation system has not been used only to decide which two soil parameters such as soil moisture and other temperature parameters use moisture, light, moisture in the air, and soil pH value. This system works as a sensor for soil water content limit value.

The most significant disadvantage of smart irrigation is the cost. These systems can be very expensive depending on the size of the property. In addition, parts of the garden should be dug to install pipes and attach them to the plumbing system in the house. It can be equal to days or weeks without using the patio. After that, the landscape should be repaired. It is better to set up an irrigation system before setting up herds or wider landscapes because some of it will have to be torn down. Groundwater is a very high-quality water supply system that is not a single area of water. Repairing the irrigation system can be far more expensive than replacing a damaged garden hose [15].

6.2 Further Improvements and Future Scope

The parameters of the agricultural environment are being monitored for a variety of reasons. But nowadays it is done to gain knowledge about the needs of the crop and soil. Here, we see irrigation from a crop farm using moisture soil sensors. The proposed soil moisture monitoring system is based on Wi-Fi, microcontroller, ESP8266 NodeMCU-ESP8266, soil moisture sensor FC-28, mobile application, and the Internet. Currently, we only use soil moisture sensors that sense the upper and lower threshold and work according to the water pump condition automatically. But anyone can use more and more to find the right state of the problem in capturing real-time measurements of that environment agriculture. Also, get treatment to overcome the problem through programming and the use of the tools involved. We are using the ESP8266 to run inexpensive, simple-to-use irrigation systems. In the future, we can securely collect all data through cloud computing and get all the information also email via the "Message in queue" service for remote transfer and mobile on the web page. The future scope of this project may include a variety of soil sensors such as temperature sensors, rain sensors, and then data collection and storage on cloud servers. This will make forecasting and analysis more accurate. It also includes creating various algorithms for appropriate data extraction for data analysis in agriculture.

References

1. Munir, M.S., Bajwa, I.S., Naeem, M.A., Ramzan, B.: Design and implementation of an IoT system for smart energy consumption and smart irrigation in tunnel farming. Energies **11**(12), 3427 (2018)
2. Sattar, H., Bajwa, I.S., Amin, R.U., et al.: An IoT-based intelligent wound monitoring system. IEEE Access **7**, 144500–144515 (2019)
3. Sarwar, B., Bajwa, I.S., Jamil, N., Ramzan, S., Sarwar, N.: An intelligent fire warning application using IoT and an adaptive neuro-fuzzy inference system. Sensors **19**(14), 3150 (2019)
4. Uddin, J., Reza, S.M.T., Newaz, Q., Uddin, J., Islam, T., Kim, J.-M.: Automated irrigation system using solar power. In: 2012 7th International Conference on Electrical and Computer Engineering, pp. 228–231 (2012). https://doi.org/10.1109/ICECE.2012.6471527
5. Abbas, Md., Mohiddin, Md., Kumar, N.: Automatic plant irrigation system. Int. J. Trend Sci. Res. Dev. **3**, 1239–1241 (2019). https://doi.org/10.31142/ijtsrd23222
6. Vanthi, M., Kavipriya, K.R., Divya Priya, D., Ambika, M.: Arduino based smart irrigation system. Ijarcce.com, 2018. [Online]. Available https://ijarcce.com/upload/2018/march-18/IJARCCE%2043.pdf
7. Nagarajapandian, M., Ram, U., Selva, G., Tamil, S.: Automatic irrigation system for sensing soil moisture content. IJIREEICE 96–98 (2015).https://doi.org/10.17148/IJIREEICE.2015.3120
8. Teame, T., Mehari, G., Mehamed, A.: Sensor based automatic irrigation management system. Int. J. Comput. Appl. Inform. Technol. **4** (2015). ISSN 2279-0764
9. Vagulabranan, R., Karthikeyan, M., Sasikala, V.: Automatic irrigation system on sensing soil moisture content. Int. Res. J. Eng. Technol. (IRJET) **3** (2016)
10. Saqib, M., Almohamad, T.A., Mehmood, R.M.: A low-cost information monitoring system for smart farming applications. Sensors **20**(8), 2367 (2020)
11. OGrady, M.J., Langton, D., O'Hare, G.M.P.: Edge computing: a tractable model for smart agriculture? Artif. Intell. Agric. **3**, 42–51 (2019)
12. Munir, M.S., Bajwa, I.S., Cheema, S.M.: An intelligent and secure smart watering system using fuzzy logic and blockchain. Comput. Electr. Eng. **77**, 109–119 (2019)
13. Safdar Malik, M., Sarwar Bajwa, I., Munawar, S.: An intelligent and secure IoT based smart watering system using fuzzy logic and blockchain. Comput. Electr. Eng. **77**(1), 109–119 (2018)
14. Sarwar, B., Bajwa, I., Ramzan, S., Ramzan, B., Kausar, M.: Design and application of fuzzy logic-based fire monitoring and warning systems for smart buildings. Symmetry **10**(11), 615 (2018)
15. Kumar, A., Kamal, K., Arshad, M.O., Mathavan, S., Vadamala, T.: Smart irrigation using low-cost moisture sensors and XBee-based communication. In: Proceedings of the Global Humanitarian Technology Conference (GHTC), San Jose, CA, USA, October 2014
16. Parameswaran, G., Sivaprasath, K.: Arduino based smart drip irrigation system using internet of things. Int. J. Eng. Sci. **6**, 5518 (2016)

An Autonomous Agricultural Robot for Plant Disease Detection

Shamim Forhad, Kazi Zakaria Tayef, Mahamudul Hasan, A. N. M. Shahebul Hasan, Md. Zahurul Islam, and Md. Riazat Kabir Shuvo

Abstract Bangladesh is the land of agriculture, known for growing varieties of crops, also the main source of employment, employing around half of the workforce. Plant diseases could have an impact on farmers' livelihoods. In order to solve this major concern, a robot is introduced that can detect plant disease using convolutional neural network (CNN), which is a sophisticated image processing algorithm. The one and only method for preventing deficits in agricultural goods quality and amount is to determine actual disease. Plant health monitoring and disease detection are essential for the long agriculture viability. The implemented robot has a solar connection for the battery package, assuring that it has a consistent source of power. The acquired image will be processed, and the classes of the plant are displayed with a local network-based website as an outcome. The website provides detected plant image with the information of disease. Due to the focus on the agriculture, renewable power, and robotics in our work, the robot is named as "AGRENOBOT."

Keywords AGRENOBOT · Agriculture · Solar panel · Robot · Arduino uno · Raspberry Pi · Image processing · CNN

1 Introduction

The agricultural land mass is more than just a source of food. Agriculture production is extremely important to Bangladesh's economy. More than half of the population is

S. Forhad (✉) · K. Zakaria Tayef · A. N. M. Shahebul Hasan · Md. Zahurul Islam
American International University Bangladesh, Dhaka 1239, Bangladesh
e-mail: mrshamimforhad@gmail.com

A. N. M. Shahebul Hasan
e-mail: shahebul@aiub.edu

Md. Riazat Kabir Shuvo
North South University, Bashundhara, Dhaka 1229, Bangladesh

M. Hasan
Bangladesh Army University of Science and Technology, Saidpur Cantonment, Saidpur, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_50

dependent on agriculture. Consequently, in the realm of agriculture, disease identification in plants is critical. The adoption of an automatic disease detection technique is advantageous in detecting a plant disease in its early stages. The diseases of the plants have the potential to harm farmers' daily life.

There are many research papers, projects, and proposed models that help in detecting plant diseases in the agricultural field. Like, Reference [1] depicts that the proposed robot is developed using deep learning algorithms like CNN to analyze the disease spread on the plant by taking a picture as a sample of the leaf. Additionally, the bot monitors the process of pesticide spraying to minimize health risks and labor work. This robot also has features like automated disease detection, and the results are sent to the corresponding mobile unit.

In [2], the proposed robot system can explore between plant rows in the fields. During sprinkling pesticides, it sends notification regularly using GSM module. The robot combines all the functions with Arduino Mega 2560. Furthermore, Raspberry Pi and camera modules are also used to detect infected plants. Besides, ultrasonic sensors and Wi-Fi modules are also used to detect objects and accessing the images of infected plants. Lastly, pumping motors and sprinkling nozzles are also used to sprinkling pesticides in the field. It can be said that the autonomous system can be an ideal solution for the farmers. Subsequent paragraphs, however, are indented.

After reviewing all these research papers, projects, and proposed plan, we came up with an updated version of robot and plant detection model to detect the disease of the plants with more accuracy and to make farming trouble-free for farmers. Our proposed agriculture robot can easily detect plant diseases with convolution neural network (CNN), implemented using image processing algorithm. In order to prevent crisis in agricultural good's quality and quantity it is important to identify the actual disease. Lack of monitoring the plant diseases, farmers can face financial loss. Farmers should have proper plan for plant disease monitoring and detection system. The autonomous robot AGRENOBOT can be a solution for this as it has a solar connectivity for supply for maximum 30 min backup with a Li-ion battery, and it doesn't need any manpower. The robot can take pictures with Raspberry Pi 4 camera module and after segmentation with VGG16 method. If there is any obstacle in the field, it can rotate up to 360°. Moreover, it can change its path while rotating. This will be a very good solution for the farmers to ensure the healthiness of the plant. AGRENOBOT has a web-based controlled system so that anyone can use it. It will just be easier to protect and maintain wide tracts of plants, along with their productivity [3, 4].

2 Methodology

The AGRENOBOT is powered by solar energy and can also be powered by charging. Here, Raspberry Pi is acting as the main processing unit and Atmega328p (Arduino uno) as a supporting micro-processing unit for robotic movement. The robot is autonomously set on an average speed of 60 rpm and sending information to the

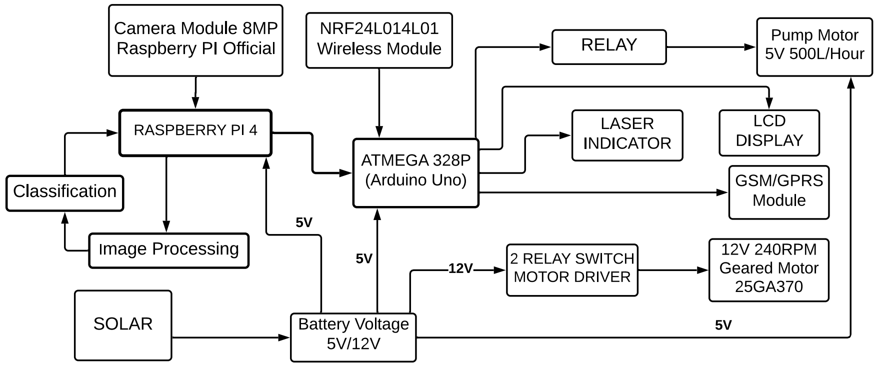


Fig. 1 System architecture

Raspberry Pi through Atmega328p. The camera is set to take pictures every second; this is matched with the rotation speed of the robot. So that the camera can take better and clear image sufficiently, with less blurriness in picture quality. The Raspberry Pi camera module will take picture and process the image according to the classification model VGG16. And after completing image processing and disease detection, the robot will send the output from Raspberry Pi to Atmega328p (Fig. 1).

In Atmega328p the robot movement process is done by conditioning. So, when the Raspberry Pi detects the disease it sends the value to Atmega328p, the process tree then takes the value and give command to break and spray to robot. The breaking period is set for 3 s and in this time the robot spray on the disease detected plant. For this spraying process, Atmega328p sends power to the pump motor, and the pump motor starts and sprays the diseased detected plant. After the spraying the robot again, start moving forward and start taking pictures. The GPS/GPRS module sends the disease information to the web server, and the local host server website shows the result and picture of the disease detected plant. The 2-relay switch gives the robot an advantage of rotating 360°. The laser indicator helps the robot to autonomously move on the agricultural field.

In Fig. 2, the diagram shows the process of power production from the solar panel to power distribution system of the robot. Here, the power is produced in solar panel. In the solar panel, the sun light hits the solar cells and from hitting the solar cells by photons electrons start flowing through the circuit in the solar panel which produce electricity in the solar panel. After producing the solar power electricity from the solar panel, we send it to the solar charger controller. The solar charger controller used to control the solar power produced by the solar. It is used to stabilize the voltage so that it does not harm the lithium-ion battery. If we don't use any stabilizer, here the unbalanced solar electricity can damage the battery. Then after stabilizing the power, we charge the battery through the power of solar. Then the power is supplied to the motor driver and other components like Raspberry Pi, Atmega328p, laser indicator, and pump. The motor driver is powered with 12 V, and for other components, the voltage is converted by DC–DC buck converter, which converts the power to 5 V, and

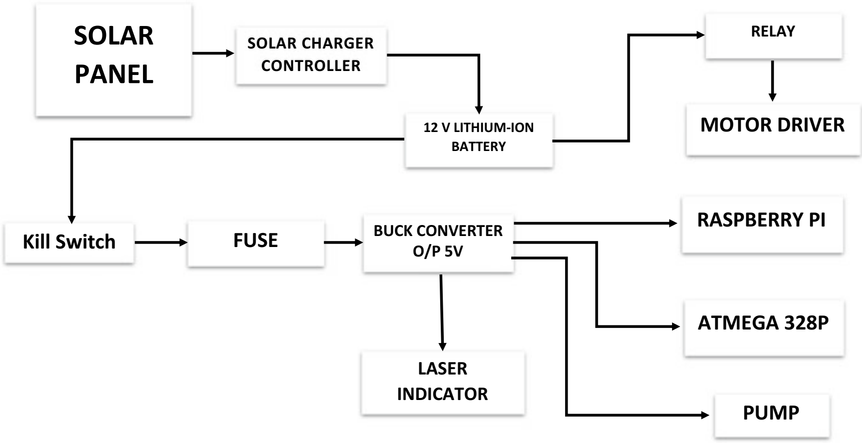


Fig. 2 Solar power system block diagram

then this is supplied to the Raspberry Pi, Atmega328p, pump, and laser indicator. That's how the solar power is distributed among the components.

2.1 Working Flowchart

The work flow diagram displayed here illustrates the workflow tree of the robot voltage supply, image processing to spray, and robot movement. After starting the robot, we can observe that it measures the battery and solar voltage. Then it returns if the battery power is sufficient. However, if it is low, it checks for a solar power voltage source, and if that is also low, it returns the alert. However, if it is sufficient, it will commence to charge the battery. It obtains the picture in the image processing phase, the taken image data is trained, and then uploads the data to the server for comparison with the pre-existing dataset. If the dataset matched with the disease affected plants, the results show affected and if not the result shows no. If the disease is detected, then it activates the pump and sprays pesticides. But if it does not find any disease then the robot moves forward autonomously or manually with the remote, we can move the remote as we want and after finishing the given command the robot start acquiring new images from the field (Fig. 3).

In Fig. 4, the autonomous and manual movement flowchart of the AGRENOBOT has been done by using Proteus 8.10. Here, The "SETUP" manifest display message, motor and light off/start status and "LOOP" indicates the directional movement of the robot according to the operation condition given in the flowchart. By following the flowchart, the robot is programmed. As we can see that the movement of the robot is contingent upon a specific set of condition and if the given conditions are satisfied the robot will execute its corresponding actions. The forward movement state, reverse

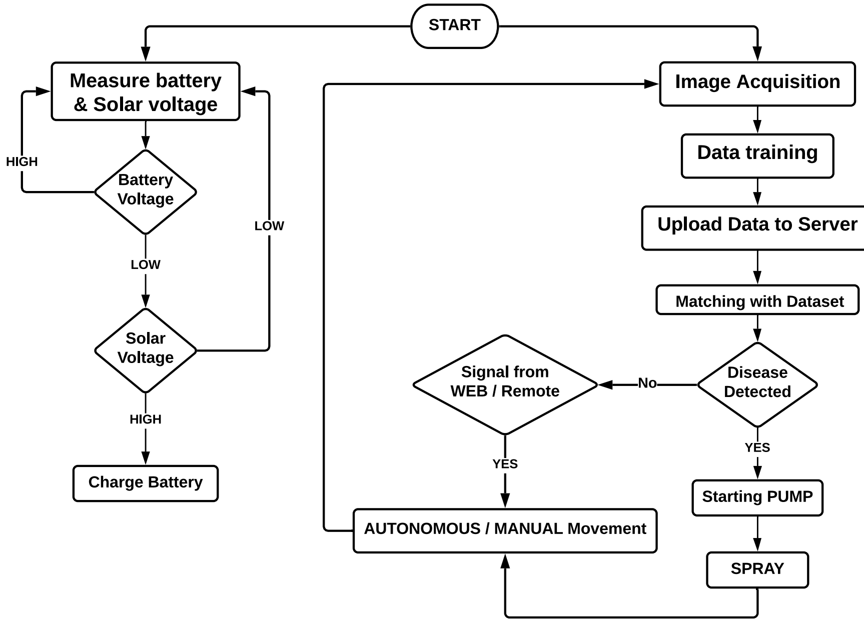


Fig. 3 Power and image processing flowchart

movement state, right movement, left movement, rotation, and breaking are defined by flowchart condition. When reverse, right, left and breaking condition is 0 and the forward moving state in 1, then the robot will move forward. Other condition of movement and breaking will follow the same method.

The spraying condition states that if the disease is detected, the robot will stop and fulfill the spraying condition. After breaking, the spraying pump will start and spray the disease detected plant for 5 s. And after fulfilling the spraying command, it will again follow the forward moving command.

2.2 *Robot Functional System of Autonomous and Manual Movement*

This (see Fig. 5) is the PCB system diagram of the robot. This PCB does the primary controlling part of the robot and giving necessary command after disease detection (Fig. 6).

The autonomous and manual movement simulation of the AGRENOBOT, which has been done by using Proteus 8.10. Here we have shown the directional movement simulation of the robot according to the operational conditions specified in the flowchart. We can see we have a display where we will receive “Welcome to

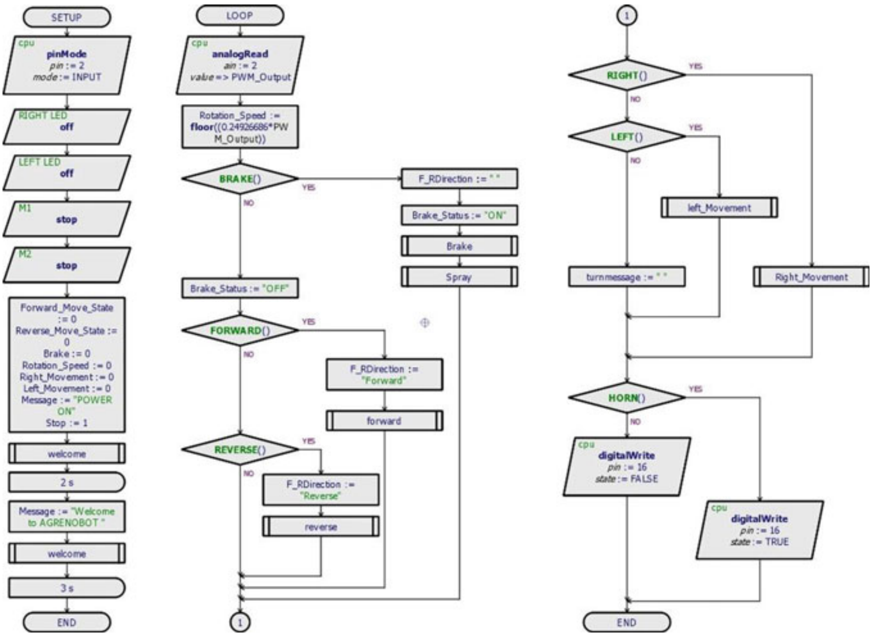


Fig. 4 Autonomous/manual movement and spraying flowchart of AGRENOBOT

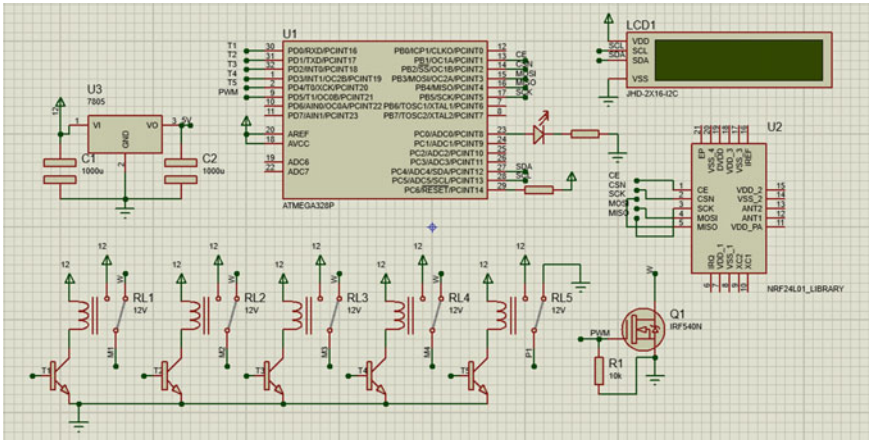


Fig. 5 Arduino uno (Atmega328p)-based robot control system

AGRENOBOT" massage after turning on the robot. And here have shown the two-motor rotational simulation for simplifying the simulation diagram. We also have two indicator which will blink when the robot will move to left or right. And we also have a horn which will beep when the robot will detect any disease in the plant.

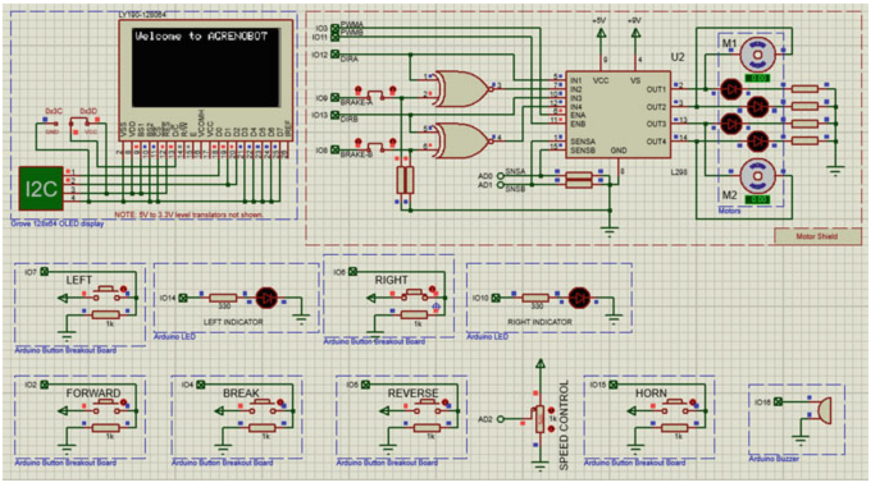


Fig. 6 Robot autonomous and manual movement simulation

When the robot detects any disease in any plant, the brake system starts working and the speed become zero. After breaking the robot spray to the plant with 10 s time, and then again, the motor start and the robot start moving forward.

2.3 Hardware Implementation of AGRENOBOT

The AGRENOBOT

This is the final robot which we have named AGRENOBOT. In this robot, we have used Atmega328p for main processing of robot controlling, which will then be directed to Raspberry Pi for command. The relays are used to control motor driver. The NRF24L01 has been used for wireless communication between robot and manual controller web/remote, and an LCD display has been used to display battery voltage, solar power, moving direction, laser, and spray on/off indication. Here, we have used 4 gear motor of 240 rpm to run smoothly of the rocky surface. Two springs have been used to balance robot body when going through uneven surface. And robot V-shape rocker body gives it an advantage to run in uneven surface in the agricultural field. As the size of the robot is smaller than the solar panel, we used the solar panel as a charging station or we can even set it on top of the robot to power it up. But the solar panel size covers more area than the robot body. We tried to find the smaller solar panel in the market but couldn't found any. The robot can operate for about 30 min. So, if we can increase the battery backup and also increase the robot size, then it can give us more energy for longer period of time. So, when the robot starts taking pictures of plants, the Raspberry Pi matches the image with the built-in VGG16 model trained image dataset. If the image is matched with the pre-existing

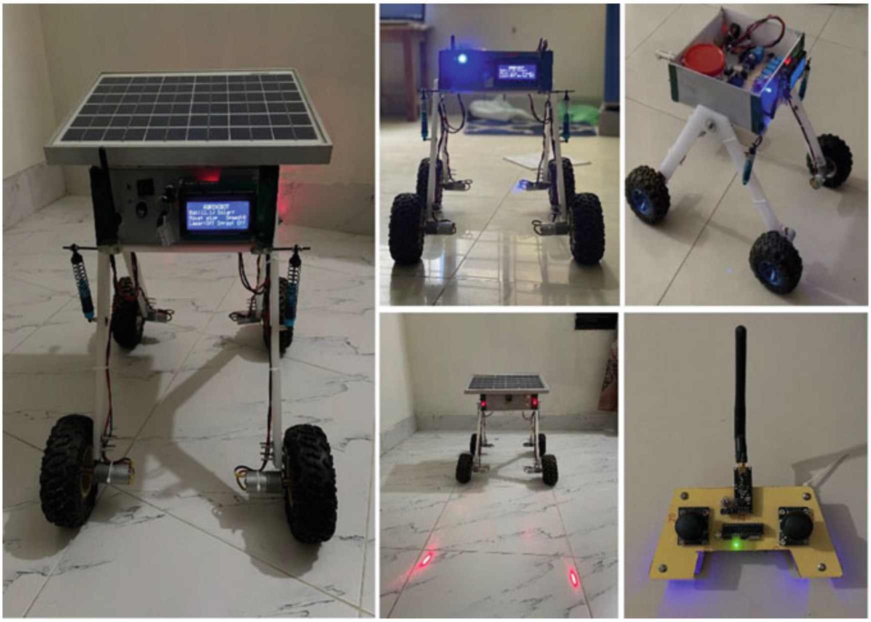


Fig. 7 Working prototype of AGRENOBOT

dataset, the result will be send to the Google Firebase Cloud to show the output in the local web system and give output as disease detected "Bacterial Spot." And also, when the disease is detected, the robot command to break through Arduino and after receiving the break signal the pump motor start and spray 1 ml pesticide to the disease detected plant. The pump motor needs 3 s to spray 1 ml pesticide, so we programmed the pump motor to spray for 3 s (Fig. 7).

Custom Remote Controller

The controller is manually built by us. The advantages of this controller are its wide range coverage and low built cost. The built cost of this remote is 5× lower than the remote controller available in market. And one more added feature is that the remote controller is rechargeable, a mobile phone lithium-ion battery is used to power the controller and can be rechargeable in very short time with a type-b charger.

2.4 Plant Disease Detection and Result Analysis Dataset

The dataset is collected from two different sources and then combined together for better result [5, 6]. The source of the dataset we have used is Kaggle (Table 1).

Table 1 Disease names and dataset samples

No	Disease name	Dataset samples
01	Bacterial Spot	2702
02	Early Blight	2920
03	Late Blight	2851
04	Tomato Mosaic	2790
05	Target Spot	2827
06	Virus Septoria Leaf Spot	2745
07	Leaf Mold	2882
08	Yellow Leaf Curl	2961
09	Spider Mites	2741
10	Healthy	2926

Implemented Image Processing Model

We did our tests using Google Colab Notebooks with a virtual GPU. We began by downloading data through Kaggle. Then, we make directories for training, validation, and testing. We then saved the above pictures to Google Drive and used them to execute the programs. Because all of the photos are JPEG files and the very first action we did was, we use TensorFlow, Keras, and ImageDataGenerator instance to decode them to RGB grid pixels. The VGG16 model was selected and utilized as the primary model during the training process. To determine performance and overfitting, we utilized accuracy and loss. We have implemented the VGG16 and archived better results. The model is implemented into the Raspberry Pi. First, we load Linux operating system into the Raspberry Pi and then imported necessary library to implement the model. And then get the dataset from desired folder and trained them with the help of the VGG16 model. And after implementing the model, we incorporated Raspberry Pi camera as an input device and Google Firebase to show the results in the local web-based system (Fig. 8).

VGG-16 has 16 convolutional layers and a remarkably consistent architecture, rendering it quite presentable. It contains basically 3 × 3 convolutions, but a lot of filters, similar to AlexNet. Transfer learning can help you attain VGG. The model is pre-trained on a dataset, the parameters are changed for improved accuracy, and the parameter values may be used. We have nine types of disease dataset in our CNN system and one healthy type dataset. Here, the robot has detected the tomato bacterial spot which is correct [8]. For classification, we have used VGG16 and VGG 16 (customized model). The model was trained using 28,345 images, and the model was validated with 24,001 images [9]. The network's input is two-dimensional picture (224, 224, 3). The first two layers have the same padding and 64 channels of 3*3 filter size. Then, after a stride (2, 2) max pool layer, two layers of convolution layers of 256 filter size and filter size (3, 3). This is followed by a stride (2, 2) max pooling layer, which is the same as the preceding layer. Following that, there are two convolution layers with filter sizes of 3 and 3 and a 256 filter. Following that, there are two sets of three convolution layers, as well as a max pool layer. Each has 512

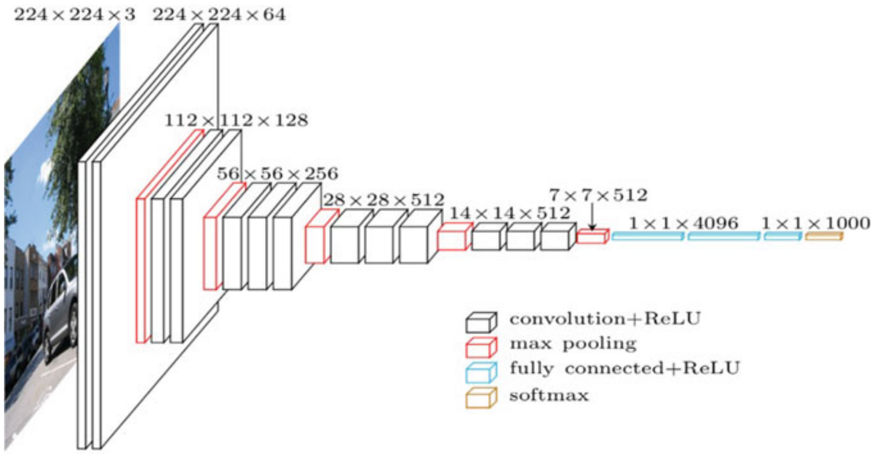


Fig. 8 Implemented VGG16 model on Google Colab [7]

filters of the same size (3, 3) and padding. This picture is then fed into a two-layer convolution stacks. The filters we employ in these convolution and max pooling layers are 3*3 instead of 11*11 in AlexNet and 7*7 in ZF-Net. It also employs 1*1 pixels in some of the layers to control the amount of input channels. After each convolution layer, a 1-pixel padding (same padding) is applied to avoid the image's spatial information from being lost. We received a (7, 7, 511) feature map after the stack of turnover and max pooling layer. This output is flattened to make it a vector (1, 5088). After this, three fully connected + ReLU layers have been created. The first layer of three connected layer receives the output of the last vector of max pooling layer and the output of first connected layer is (1, 4096) vector, and the second layer likewise outputs the same size vector (1, 4096), while the third layer produces the 1000-channel output for 1000 classification [10] (Fig. 9).

Disease detection result

Here, we have shown the traditional or existing VGG16 model result and the customized VGG16 result used in the project to identify plant disease detection more accurately. In this project, we used accuracy, precision, recall, and F1 score to determine the performance of the deep learning-based plant disease detector. The accuracy (A) is the percentage of presently categorized forecasts that are correct. Precision (P) is the fraction of positive outcomes that were ultimately right. The proportion of verifiable positives that were appropriate implementation is measured by recall (R), and the F1-score is derived by taking the harmonic mean of accuracy (Table 2).

IOT-based web result of disease detection and control system

Here in Fig. 10, we can see the web-based live feedback. With this web-based live feedback, we can both control the robot manually and see the output of the detected

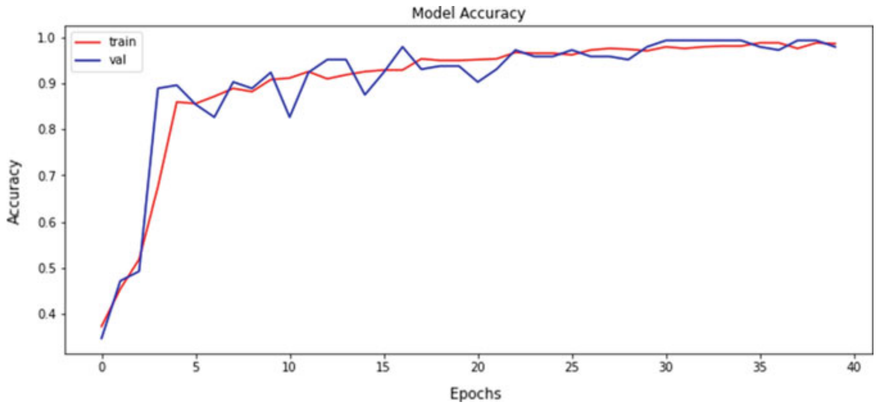


Fig. 9 Implemented VGG16 (customized) model accuracy

disease after scanning of the image. The robot takes the picture and sends it to the Raspberry Pi for image processing. The Raspberry Pi checks which category the picture belongs to by classifying the images. Then when the disease is detected or healthy plant is detected, the Raspberry Pi sends the taken picture and result to the live web script, using the firebase real-time database, and viewers can gain safe access to the database from robot's Raspberry Pi code. Data is saved locally, and real-time activities continue to display updated results even when the website is offline, offering every user an interactive and dynamic response. That's how through firebase real-time database the live web script shows the final result of the image and detected plant.

3 Conclusion

The AGRENOBOT has been constructed to benefit farmers by significantly improving their productivity and efficiency. Farmers labor all day and encounter several challenges; therefore, this system would assist them in reducing their work-load while also assisting them in increasing their output through regular monitoring plant diseases. The identification of disease benefits the farmer in producing healthy crops. Robot's pesticide sprinkling function eliminates the need for the farmer to manually apply pesticides. In near future, the demand of the agriculture robot will increase surely as it will save a lot of money of the farmers. It will play a significant role in our country's economy. This project's eventual goal is to make a cost friendly robot with higher work done and less return loss and 100% accuracy on plant disease detection using customized image processing model VGG16. As this project is eco-friendly, so it is not harmful to the environment. As the world runs so rapidly nowadays, agriculture is essential in our daily lives. This project is one of

Table 2 Accuracy, precision, recall, and F1-score of VGG16 and VGG19

Plant	Disease name	VGG16				VGG16 (Customized)			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Tomato	Bacterial Spot	0.77	0.79	0.81	0.85	0.97	0.92	0.95	0.97
	Early Blight	0.77	0.79	0.82	0.85	0.97	0.92	0.96	0.97
	Late Blight	0.77	0.79	0.82	0.82	0.97	0.92	0.95	0.98
	Mosaic Virus	0.76	0.76	0.81	0.83	0.97	0.91	0.96	0.98
	Target Spot	0.74	0.77	0.8	0.8	0.96	0.92	0.95	0.97
	Septoria Leaf Spot	0.75	0.75	0.81	0.82	0.97	0.91	0.96	0.98
	Leaf Mold	0.72	0.72	0.78	0.82	0.96	0.92	0.96	0.98
	Yellow Leaf Curl	0.76	0.73	0.81	0.81	0.97	0.92	0.95	0.96
	Spider Mites	0.75	0.71	0.82	0.82	0.97	0.93	0.97	0.99
	Healthy	0.72	0.64	0.74	0.79	0.95	0.7	0.79	0.89

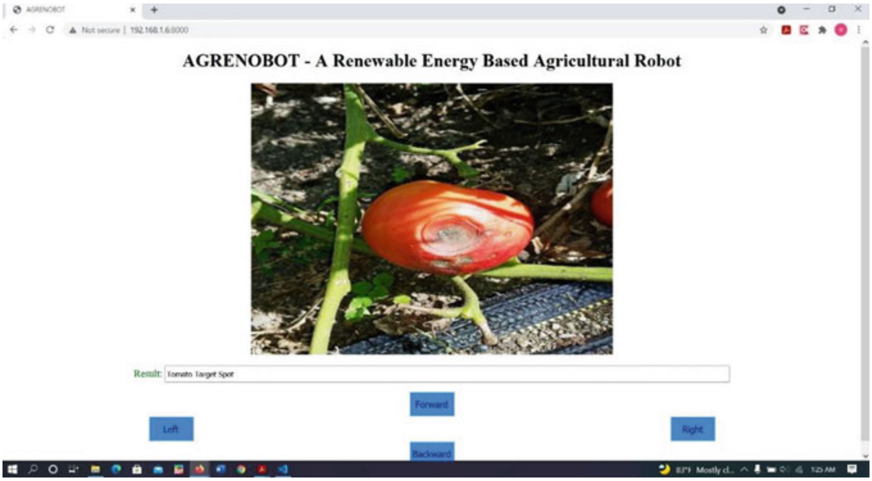


Fig. 10 Web-based detected disease image and result

the best options now and, in the future, it will help to develop the agriculture system and contribute more in the future precision agriculture.

References

1. Shetty, S., Bhosale, R., Naik, R., Mane, R., Lokare, R., Patil, S.: Agribot-system for plant disease detection and pesticide sprinkling (May 7, 2021)
2. Bhoite, P.S., Gurav, N.U., Nagarkar, O.K., Chaudhari, R.R.: SEKTROO: smart automated pesticide sprinkling bot. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1420–1423 (2019). https://doi.org/10.1109/ICOEI.2019.8862729
3. Sharif Ullah Al-Mamun, G.M., et al.: Performance analysis of multipurpose AGROBOT. In: 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECONECE), pp. 1–4 (2019). https://doi.org/10.1109/WIECONECE48653.2019.9019906
4. Ranjitha, B., Nikhitha, M.N., Aruna, K., Afreen, Murthy, B.T.V.: Solar powered autonomous multipurpose agricultural robot using Bluetooth/Android App. In: 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 872–877 (2019)
5. Tomato | Kaggle, 14 Sept. 2018. Accessed on: 6 Oct. 2021. [Online]. Available https://www.kaggle.com/noulam/tomato
6. Tomato leaf disease detection | Kaggle, 7 Aug. 2016. Accessed on 10 Oct. 2021. [Online]. Available https://www.kaggle.com/kaustubhb999/tomatoleaf
7. Vgg19—Towards Data Science, Sept. 2018. Accessed on 22 Aug. 2021. [Online]. Available https://towardsdatascience.com/tagged/vgg19
8. VGG-16 | CNN model, 26 Feb. 2020. Accessed on 5 July 2021. [Online]. Available https://www.geeksforgeeks.org/vgg-16-cnn-model/

9. Karthik, R., Hariharan, M., Anand, S., et al.: Attention embedded residual CNN for disease detection in tomato leaves. Appl. Soft Comput. J. (2019). https://doi.org/10.1016/j.asoc.2019.105933
10. Step by step VGG16 implementation in Keras for beginners, 6 Aug. 2019. Accessed on 13 June 2021. [Online]. Available https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-forbeginners-a833c686ae6

Design and Development of a Low-cost IoT-Based Water Quality Monitoring System

Sultanus Salehin, Tahseen Asma Meem, Akib Jayed Islam, and Nasim Al Islam

Abstract As a consequence of rising urbanization and industrial growth, water contamination and degradation are developing at an alarming rate. The water scarcity around the world necessitates a long-term strategy to make the most of it. Traditionally, water quality is assessed by collecting water samples by hand and then testing and analyzing them in a laboratory setting. This paper examines how cutting-edge technology, such as the Internet of Things (IoT), can provide a sustainable and cost-effective method of monitoring multiple water parameters in real time. The proposed system was used to calculate the turbidity, TDS, pH level, and temperature of 30 different water samples with success. The turbidity level was measured in nephelometric turbidity units (NTU) and then transmitted via wireless fidelity (Wi-Fi) networks to an Internet of Things—cloud computing platform, where it could be viewed using an Android smartphone or PC. The experiments demonstrated that the monitoring system was capable of continuously monitoring the pH level, total dissolved solids (TDS), and temperature of water from various sources at different times, thereby providing safe water for industrial, agricultural, and commercial purposes. The cost and complexity of implementation are minimal due to the use of sensors and the Arduino Nano microcontroller, making it simple to validate the efficacy of the built system.

Keywords Water quality monitoring · The pH level · Temperature · Turbidity · IoT · Sensors · TDS.

S. Salehin and T. Asma Meem: Authors contributed equally to this paper.

S. Salehin (✉)
Islamic University of Technology, Gazipur, Bangladesh
e-mail: sultanussalehin@iut-dhaka.edu

T. A. Meem · A. J. Islam · N. Al Islam
American International University-Bangladesh (AIUB), Dhaka 1229, Bangladesh

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_51

1 Introduction

The four basic constituents of our nature are the atmosphere, meaning air, lithosphere that includes rocks and soil, hydrosphere or water, and living beings that create biosphere. Water and oxygen are essential for life on Earth; in fact, biologists believe that water is the source of life [1]. Fresh natural water is an essential source of survival, and its consumption is fast growing as natural freshwater resources become scarce. Improper water resource management has resulted in severe water shortages in several parts of the world. Over the last few decades, an exponential increase in population growth, urbanization, and industrialization has resulted in ever-increasing water pollution in Bangladesh and many other developing countries throughout the world. In Bangladesh, where river water is in short supply, it is widely used for home, agricultural, and industrial purposes. Over the past 40 years, severe environmental contamination issues have occurred in the waters encompassing Dhaka and the Karnaphuli River in Port City, Chattogram. According to current data, several urban rivers including the Teesta, Korotoa, Pashur, Rupsha, and Padma are in grave danger. Waterborne pollutants wreaked havoc on riverbeds downstream [2]. Human health, as well as the ecological balance among other species, is impacted as a result of water contamination.

According to one study, approximately 1.8 B individuals in different parts of the globe could encounter complete water dearth by 2025 [3]. Still pollution, which is mostly produced by power generation, industrial production, mining, and other aspects, is a major contributor to the degradation of water bodies, exposing individuals to cholera, diarrhea, dysentery, hepatitis A, and other waterborne diseases. Diarrhea claims the lives of an estimated 842,000 individuals each year [4]. As a result, it is the time to develop an autonomous monitoring system to ensure water quality balance [4].

The number of surveys has been carried out to determine the water constituents of an environment using just glass, scoops, and bottles. Water samples are gathered, and the findings are examined in a laboratory to identify the water constituents, which is a time-consuming and traditional technique. World research has demonstrated the IoT, i.e., Internet of Things, to be a cutting-edge technical phenomenon due to its versatility in collecting, monitoring, and analyzing data gathered from even the most remote sites [5, 6]. The ultimate goal of all recent IoT-related research in this field remained to bring about an efficacious, low-cost, instantaneous system to observe the quality of water that coalesced wireless sensor networks and the Internet of Things together [7]. In this study, we proposed a water quality monitoring system that consists of a MCU and fundamental sensors, is small-scale, and is highly beneficial for monitoring pH, turbidity, TDS, water flow with water temperature, and then uploading the sensor data to the cloud.

The remainder of this research work is structured in the succeeding sections: Sect. 2 summarizes the project's related works, Sect. 3 describes the design procedure and hardware implementation of the proposed system, Sect. 4 shows the results and discussions, and Sect. 5 summarizes the project's future goals and conclusion.

2 Related Works

In this current era, water quality monitoring applications are particularly incorporating IoT, a system where almost all the devices can work autonomously and interact with other devices without any human involvement. In order to construct a high-quality model, we looked at a variety of different methodologies developed by researchers for this type of system. Authors have developed distinct methods for determining water quality by examining characteristics such as turbidity, pH, temperature, conductivity, and so on. Theofanis Lambrou et al. [8] worked with a water-level controlling system focusing on multiple key characteristics, i.e., mobility, portability, cost-effectiveness, and reliability, and they employed a sensor array along with multiple microsystems in order to carry out-signal conditioning and processing. The wireless system made it possible to remotely process the data and reduce the costs of installation significantly. However, the proposed method could not cover all types of contamination. Saravanan et al. [9] found that performance parameters and working procedures changed with time. They suggested a cutting-edge SCADA system which stands for supervisory control and data acquisition for real-time water quality monitoring that interacts with the Internet of Things (IoT), which was installed in a Tamil Nadu metropolis. Arduino Atmega 368 using GSM modules was used to identify the contamination level of water as well as pipeline leakage. The whole system was made cost-effective upon calibration with additional sensors. However, in SCADA systems, which run 24×7 in factories, minimizing energy consumption of low-power sensors is also a challenging issue.

In the research work done by Vasudevan and Balraj Baskaran [10], a novel water quality monitoring system was proposed which turned out to be an efficient one in the process of eradicating pollution control trends. They built an unmanned surface vehicle (USV) to gather information about quality parameters, chemical index in particular, analyze, and create characteristic curves of each of them about the water body under experiment. The whole process of data collection and tracking of quality parameters and assessment was done in real time as the data were transferred immediately to the cloud.

Cloete et al. [11] established a water quality monitoring system that avoided wired communication to disseminate the collected data and could display the acquired data via an UI (User Interface). It had an automated warning notification option to alert the users if any parameter goes beyond the safety level but had no way to keep track of the measurement history due to the unavailability of a data logging system. Mohammad Salah Uddin Chowdhury et al. described a water quality monitoring system which is WSN based with inter- and intra-node communication in their publication "IoT Based Real-time River Water Quality Monitoring System." On a server PC, data visualization is performed using Spark streaming analysis using neural network models, a belief rule-based system, and Spark MLlib. This work proved to be particularly beneficial in the context of Bangladesh since it concentrated on high mobility, high frequency, and low power. However, the authors concentrated only on river water characteristics (temperature, pH, and turbidity), leaving additional factors such as dissolved oxygen (DO), chemical oxygen demand (COD), total dissolved solid (TDS), and cost-effectiveness and resilience to be explored later [12].

Traditional methods for measuring water quality, such as the time-consuming manual laboratory examination process and statistical inferences [13], are time-consuming, expensive, and inefficient. Water quality is influenced by biological, chemical, and physical forces in nature. Because a single metric cannot adequately quantify water quality, the water quality index (WQI) was devised to do so. Several WQI calculation techniques were developed, with the majority focusing on identifying contamination occurrences. Sandia National Laboratories created Canary, one of the earliest such systems, with funding from the Environmental Protection Agency (EPA). Greater Cincinnati Water Works is now using it (GCWW). It surpasses previous systems in several critical areas, including algorithm lucidity, the capacity to incorporate operational info directly into its event identification component, and the ability to possess concentrated processing on an independent computer system while buttressing numerous sensors. Elad Salomons' OptiEDS is another device that assists in the real-time observation of abnormal water-quality situations including water quality monitoring and detection of water contamination. Another instrument that can identify anomalous behavior in water quality tests is the Bluebox. Because it first normalizes the data before calculating the distance between the variables in each data point and uses distances to plot the frequency curves to represent the result, it gives a valid outcome even if some of the variables are absent. It is, however, rather exorbitant in terms of money, costing up to $92,500, and it lacks the ability to incorporate operational info directly into event detection. Another system developed by the Hach Company is Event Monitor, having an investigatory capacity to detect events, tune itself mechanically, and determine what accounts for an anomaly inside the system [14].

3 Design Procedure and Hardware Implementation of the Proposed System

The need for a better methodology for water quality monitoring system (WQMS) has always been on the increase due to the challenge of the limitation of water resources in a developing country like Bangladesh. In general, numerous parameters must be measured in order to perform a water quality analysis. So, keeping the focus on vital water parameters, we proposed our WQMS model into two parts—hardware implementation and software part. Proteus Professional 8.0 was used for simulation purposes after which we proceeded for hardware implementation. Figures 1 and 2 depict the proposed WQM system's block design and flow chart. All sensors are linked to the central MCU, including the temperature sensor, pH sensor, TDS sensor, turbidity sensor, and water flow sensor, and these sensors feed the MCU with necessary data from the water reservoir. The information is sent to the cloud immediately through a Wi-Fi module, where it may be seen at any time from an Android device or a PC.

The Arduino Nano is a microcontroller board with 8 analog pins, 14 digital pins, 6 power pins, and 2 reset pins that may be used on a breadboard. It is programmed using the Arduino IDE. This and the Arduino UNO are almost similar boards in terms of capability. It operates at 5 V, although the input voltage can be between 7 and 12 V. Because the Arduino Nano's maximum current rating is 40 mA, any load attached to its pins should not use more than 40 mA for best results [15].

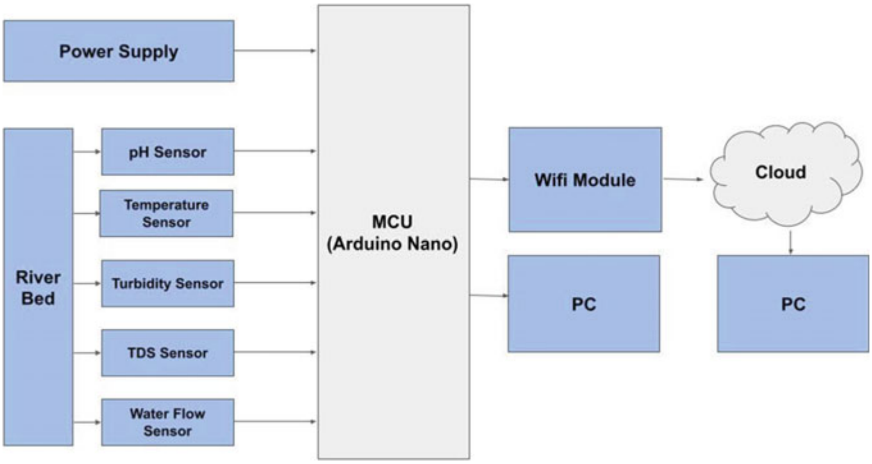


Fig. 1 Block diagram depiction of the proposed WQM system

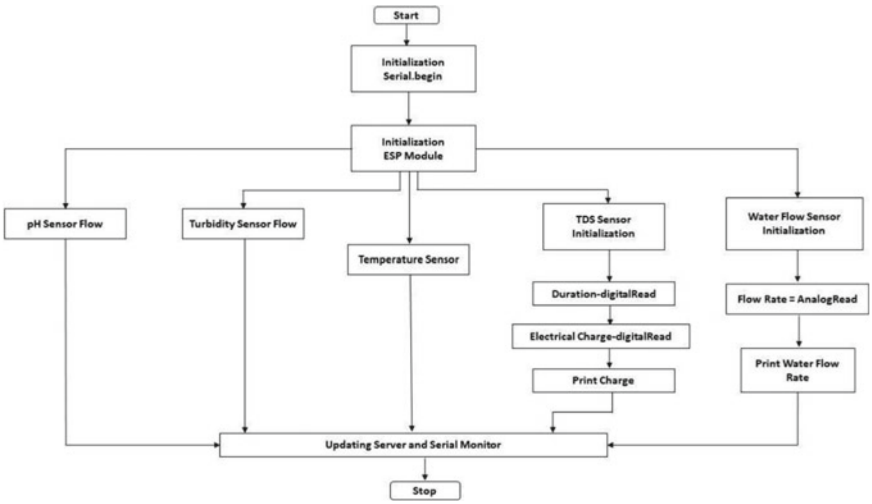


Fig. 2 Proposed system

The ESP8266 ESP-01 Wi-Fi module is a self-sufficient System on a Chip (SOC) that allows a microcontroller to connect to a wireless network and manages its inputs and outputs without the need for a microcontroller, since it functions similarly to a tiny computer. As a result, it may be set up to connect to a Wi-Fi network. Also, it can operate as a microcontroller.

The pH scale, ranging from 0 to 14, is used to represent a solution's alkalinity or acidity, which is calculated by the quantity of hydrogen [H] + ions available. The potential difference between two electrodes: a reference node and a glass electrode sensitive to hydrogen ions may be used to compute PH levels in this module. This Arduino-compatible sensor may be used to assess water quality and for aquaculture [16].

The water turbidity sensor for Arduino determines the standard of water by measuring turbidity. It calculates the rate/percentage of light scattering and transmission, which is proportional to the amount of total suspended solids (TSS) in the water, to point out suspended particles in water. The TSS grows in lockstep with the amount of liquid turbidity [17]. A water flow sensor is made consisting of a water rotor, a plastic valve body, and a hall-effect sensor. When water passes through the rotor, it rolls. Its speed is determined by the flow velocity. The Hall effect sensor then produces the proper pulse signal [17]. Despite the fact that the sensor works well up to 125 °C, it is recommended that the temperature be kept below 100 °C due to the PVC jacketing on the cable. Only a wire (and ground) is required to connect the DS18B20 sensor to a microprocessor. Many DS18B20s can be utilized on the same 1-wire bus system at the same time since each DS18B20 has a unique silicon serial number, allowing temperature sensors to be put in a variety of places within the same project. The analog TDS sensor/meter for Arduino is an Arduino-compatible TDS kit that may be used to calculate the TDS value of water in order to evaluate its purity. The total dissolved solids (TDS) in one liter of water are measured in milligrams. More soluble particles dissolve in the water when the TDS level is high, making the water less pure [16]. Our proposed system includes pH sensor, water turbidity sensor, water flow sensor, waterproof DS18B20 digital temperature sensor, and analog TDS sensor. As the principal processing module, we employed Arduino Nano, which consumes low power, is small and compact in size and one of the most significant parts of the WQMS. To measure the acidity or the alkalinity, we used a pH sensor while a water turbidity sensor was used to detect the suspended particle which enables the user to determine the light scattering and the transmitting rate of the available water. The use of a temperature sensor was intended to observe the real-time temperature of the water resource along with the surrounding environment in order to verify that the water quality was sound. Furthermore, by estimating the total dissolved solids in the water, the TDS sensor aids in determining the purity of the water supply. The software experimental set up of our suggested system is illustrated in the software simulations shown in the following Figs. 3 and 4.

The microcontroller, which includes a SoC analog-to-digital converter that converts the analog signals received by the sensors into digital signals for in-depth analysis, is used for data sensing and control from sensors. To receive this analog output, MCU's analog pins and the sensor's analog output must be attached. The

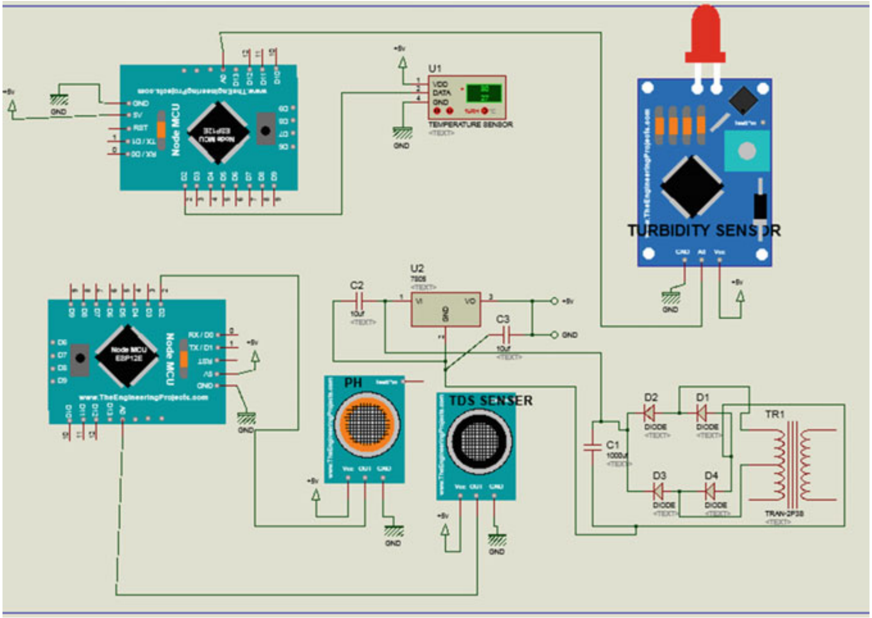


Fig. 3 Schematic diagram of software simulation

processed data is updated to the central server using a low-cost ESP8266 ESP-01 Wi-Fi module. A Wi-Fi chip with a full TCP/IP stack, as well as a microcontroller chip and Tx and Rx serial transceiver pins that send and receive data, modify wireless module framework, and update serial inquiry instructions, are included.

4 Results and Discussion

The sensor network of the MCU of the experimental set up is capable of capturing sample data from the water storage reservoir every 30 s. The Arduino IDE serial display presents the sample parameter outputs. The Wi-Fi module is used for real-time monitoring with different parameters. With the help of a pH sensor, the voltage of the water is continuously being updated with corresponding values. The water's pH is proportional to the measured voltage of the water according to the Nernst equation,

$$E = E^o + (RT/zF)\text{pH}$$

Here, cell potential is denoted by E while E^o refers to the cell potential in conventional pressure and temperature state, universal gas constant is denoted by R, number of electric moles is denoted by z, F refers to the constant Faraday, and T is the

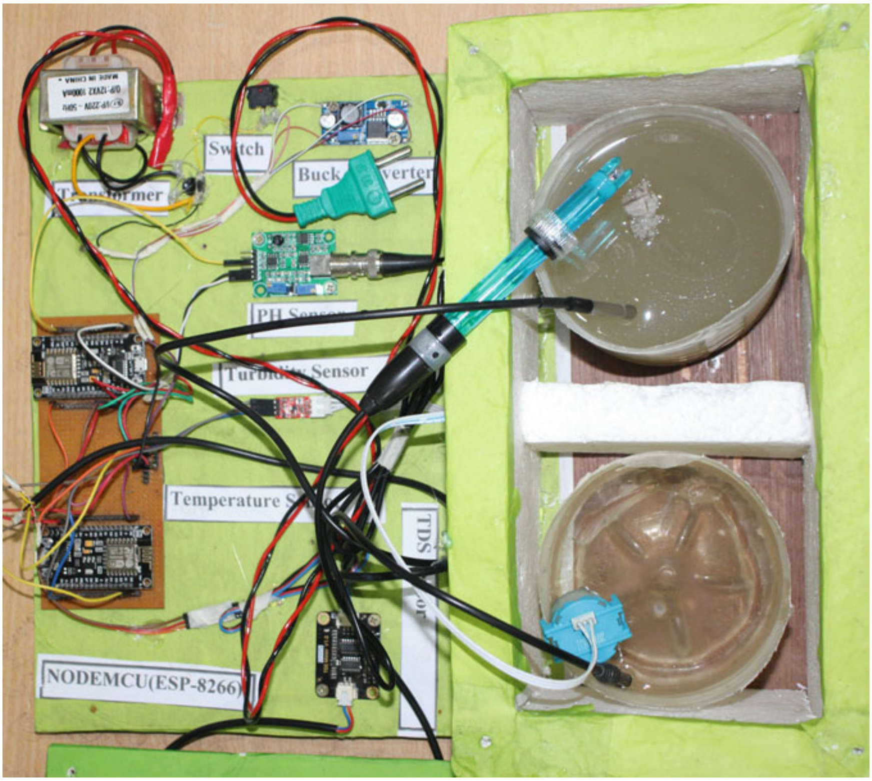


Fig. 4 Experimental setup

temperature. In Arduino IDE's serial monitor, Figs. 5 and 6 display the output of the pH sensors. In Figs. 5 and 6, we observed 30 samples from 30 different resources at different time periods and the result of the pH levels quite depicts the actual scenario apart from some exceptional occasions. The corresponding values of total dissolved solids (TDS) as well as turbidity are calculated using gravity: TDS sensor and turbidity sensor, respectively.

Figures 7 and 8, respectively, give us the idea of TDS values and turbidity values of 30 unique resources at different time periods. The turbidity is measured in NTUs, or nephelometric turbidity units, and the turbidity sensor outputs are examined and updated on the Arduino IDE serial monitor.

$$y = -1120.4x^2 + 5742.3 - 4352.9$$

According to the preceding equation, the water turbidity is reciprocal to the voltage of the water, where x is the value of voltage and y is the turbidity value.

The pH level of drinking water recognized for consumption globally must be between 6.5 and 8.5, but our suggested model demonstrates 21 samples of water

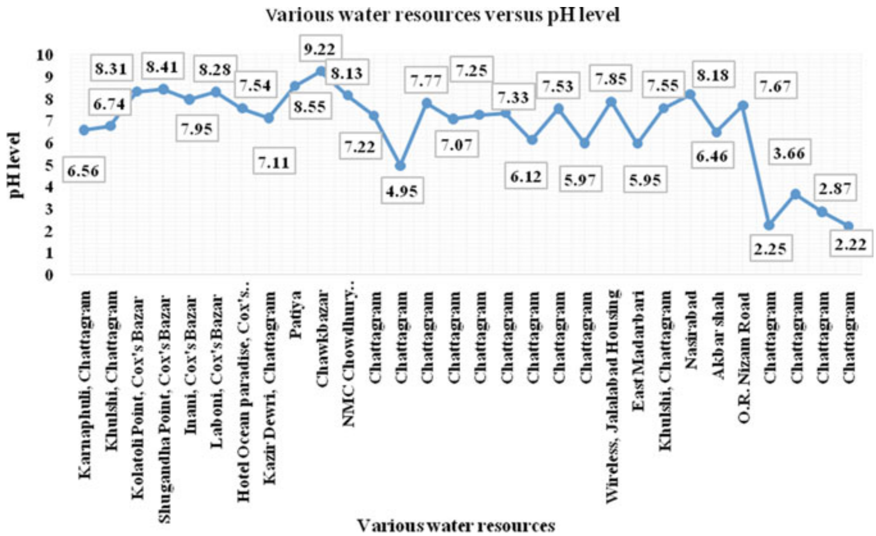


Fig. 5 Various water resources versus corresponding pH level

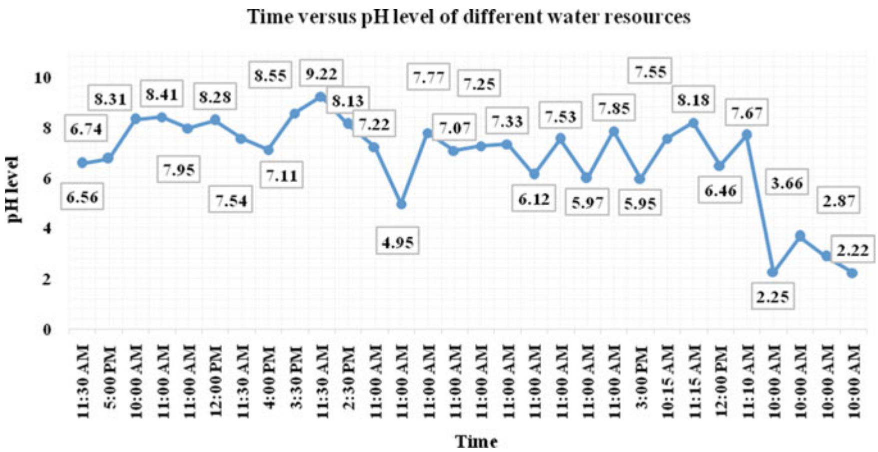


Fig. 6 pH level of different water resources at different time period

from 30 different water resources with pH levels between 6.5 and 8.5 and can be considered safe to drink. These 21 samples of water were collected as follows: sea water from Cox's Bazar (Kolatoli Point, Shugandha Point, Inani, and Laboni) and sea water, tap water, and deep groundwater from Khulshi, Chattagram, bottled water of companies (Acme, Jibon, PRAN, Spa Fresh, MUM) in Chattagram and swimming pool water from Hotel Ocean Paradise, Cox's Bazar. From Nasirabad, Akbar Shah Road, O. R. Nizam Road, we have fetched deep groundwater as well. TDS values of less than 300 ppm are preferred and appropriate for drinking without filtration.

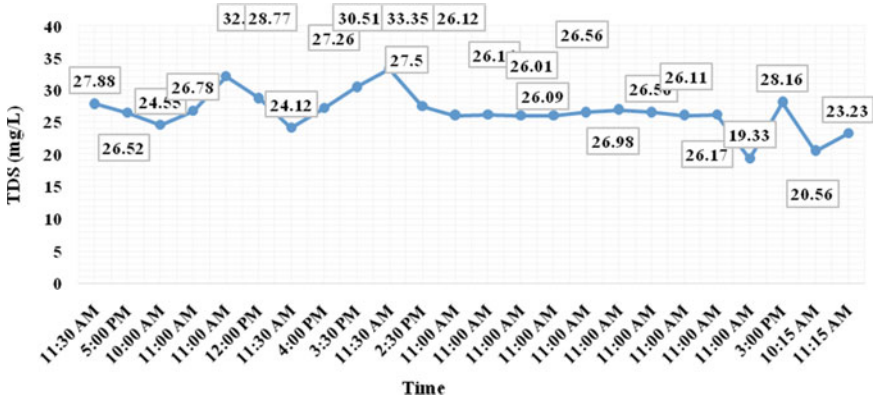


Fig. 7 Total dissolved solids in different resources at different time

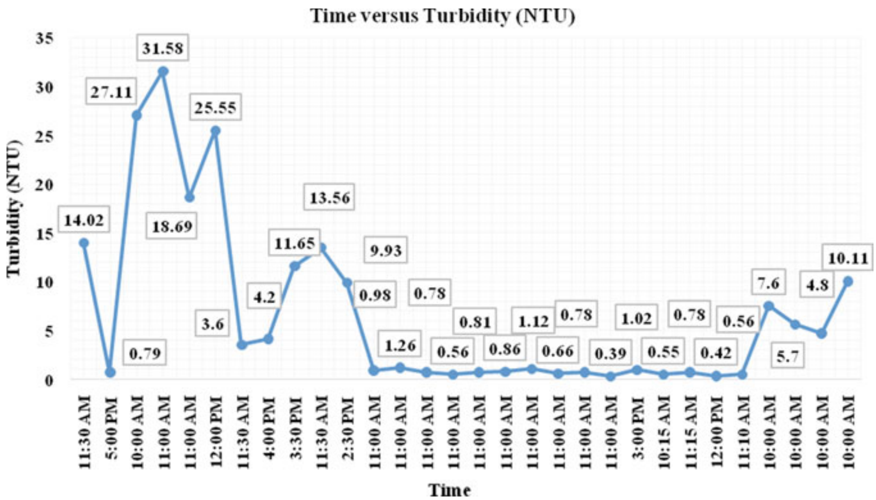


Fig. 8 Turbidity in NTU of different water resources at different time period

Bottled water of companies (Aquafina, Acme, Jibon, PRAN, Fresh, Kinley, MUM) has TDS levels below 300 ppm, making them safe to consume without filtration. The turbidity level in drinking water that is deemed to be more desirable and acceptable for consumption must be less than 1 NTU. Our suggested model demonstrates that out of 30 distinct sources, 13 samples of water had turbidity levels below 1 NTU, making them safe to consume.

The waterproof DS18B20 digital temperature sensor is employed to determine the accuracy of pH and turbidity measurements by measuring the water surface temperature. The temperature of the environment is determined using the temperature sensor since the pH sensor and turbidity sensor will yield proper data under a specified atmospheric state. Figure 9 represents the corresponding temperature value of the same water resources at distinct time periods.

The temperature of drinking water that is acceptable for ingestion around the world must be fewer than 25 degrees Celsius. Our suggested approach reveals that out of 30 different resources, 6 samples of water have a temperature of less than 25 degrees Celsius and are safe to consume.

Table 1 shows the cost analysis of the proposed WQM system. The prototype model is less expensive than other available systems, as shown in the table. All of the equipment's costs are current, since they were purchased in July of 2021. Aside from temperature sensors, the majority of sensors have a lifespan of 1–3 years; however, a temperature sensor may last up to 10 years. All of the sensors have high sensitivity and accuracy in terms of success rate.

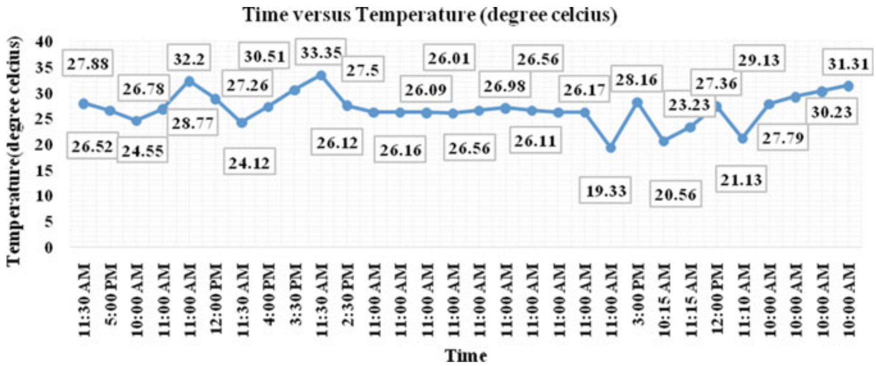


Fig. 9 Temperature versus time curve of several water resources

Table 1 Cost analysis of WQM system

Name of the component	Model number (Quantity)	Price per unit in Bangladesh taka
Arduino Nano V3	ARD-00089 (1)	398.03
pH sensor with module	SEN-00239 (1)	2499.22
Water turbidity sensor	SEN-00205 (1)	780.18
Temperature Sensor	SEN-00072 (1)	650.53
Name of the component	Model number (Quantity)	Price per unit in Bangladesh taka
ESP8266 ESP-01	WIR-00064 (1)	295.23
G1/2" Water flow sensor	SEN-00043 (1)	479.61
Gravity: Analog TDS Sensor/meter for Arduino	SEN-00222 (1)	1484.68
Total costing		6,587.48

5 Future Work and Conclusion

The proposed technique for monitoring water parameters has yielded promising preliminary findings. Despite the fact that the water monitoring system was developed to collect samples from multiple water sources in Chittagong as well as Cox's Bazar, Bangladesh, the conclusion can be drawn that the system can be used for personal usage as well. This proposed cutting-edge IOT technology can be used to keep track of the water standard of any water source in any part of the planet in the future. In the future, the system can be made more robust and compact by complex PCB design [17–20] where we can add an additional security system as well as a live data visualization system. The inclusion of an Android [5, 18, 19] app can be a huge change in the entire WQM system. Finally, it can be concluded that with this proposed paradigm, the goal of ensuring safe drinking water for all can be achieved.

References

1. Hasan, Md.K., Shahriar, A., Ullah Jim, K.: Water pollution in Bangladesh and its impact on public health. Heliyon **5**(8), e02145 (2019)
2. Uddin, Md.J., Jeong, Y.-K.: Urban river pollution in Bangladesh during last 40 years: potential public health and ecological risk, present policy, and future prospects toward smart water management. Heliyon **7**(2), e06107 (2021)
3. WHO: Drinking-water fact sheets. Retrieved from World Health Organization (2017a). http://www.who.int/mediacentre/factsheets/fs391/en/. Last accessed 21 09 2021
4. Daigavane, V.V., Gaikwad, M.A.: Water quality monitoring system based on IoT. Adv. Wireless Mob. Commun. **10**(5), 1107–1116 (2017)
5. Islam, A.J., Farhad, M.M., Alam, S.S., Chakraborty, S., Hasan, M.M., Nesar, M.S.B.: Design, development and performance analysis of a low-cost health-care monitoring system using an android application. In: 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), pp. 401–406. IEEE (2018)
6. Islam, A.J., Hamid, M.K., Alam, S.S., Hasan, M.M., Hoque, M.A., Rafid, M.N.M.: A greener approach to harvest energy using piezo-speed breaker. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–4. IEEE, (2019)
7. Alam, S.S., Islam, A.J., Hasan, M.M., Rafid, M.N.M., Chakma, N., Imtiaz, M.N.: Design and development of a low-cost IoT based environmental pollution monitoring system. In: 2018 4th International Conference on Electrical Engineering and Information and Communication Technology (iCEEiCT), pp. 652–656. IEEE (2018)
8. Lambrou, T.P., Anastasiou, C.C., Panayiotou, C.G., Polycarpou, M.M.: A low-cost sensor network for real-time monitoring and contamination detection in drinking water distribution systems. IEEE Sens. J. **14**(8), 2765–2772 (2014)
9. Saravanan, K., Anusuya, E., Kumar, R., Hoang Son, L.: Real-time water quality monitoring using Internet of Things in SCADA. Environ Monit. Ass. **190**(9), 1–16 (2018)
10. Vasudevan, S.K., Baskaran, B.: An improved real-time water quality monitoring embedded system with IoT on unmanned surface vehicle. Eco. Inform. **65**, 101421 (2021)
11. Cloete, N.A., Malekian, R., Nair, L.: Design of smart sensors for real-time water quality monitoring. IEEE Access **4**, 3975–3990 (2016)
12. Chowdury, M.S.U., Emran, T.B., Ghosh, S., Pathak, A., Alam, M.M., Absar, N., Andersson, K., Shahadat Hossain, M.: IoT based real-time river water quality monitoring system. Proced. Comput. Sci. **155,** 161–168 (2019)

13. Gazzaz, N.M., Yusoff, M.K., Aris, A.Z., Juahir, H., Ramli, M.F.: Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Marine Pollut. Bullet. **64**(11), 2409–2420 (2012)
14. Ahmed, U., Mumtaz, R., Anwar, H., Mumtaz, S., Mustafa Qamar, A.: Water quality monitoring: from conventional to emerging technologies. Water Supply **20**(1), 28–45 (2020)
15. [Online]. Available at: www.theengineeringprojects.com/2018/06/introduction-to-arduino-nano.html. Last accessed 21 Sep 2021
16. [Online]. Available at: www.techshopbd.com. Last accessed 21 Sep 2021
17. Islam, A.J., Alam, S.S., Ahammad, K.T., Nadim, F.K., Barua, B.: Design, kinematic and performance evaluation of a dual arm bomb disposal robot. In: 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), pp. 1–6. IEEE (2017).
18. Alam, S.S, Islam, A.J., Hasan, M.M., Farhad, M.M.: Design and Implementation of an embedded system to observe the atmospheric condition using a helium balloon. In: 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), pp. 242–246. IEEE (2018)
19. Alam, S.S, Islam, A.J., Hasan, M.M., Imtiaz, M.N.: A low-cost GPS based application for navigating shallow waters. In: 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), pp. 92–97. IEEE (2018)
20. Alam, S.S, Islam, A.J., Hasan, M.M., Nadim, F.K.: Energy harvesting from dual tone RF signal using a double stage voltage rectifier for wireless power transmission. In: 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), pp. 757–762. IEEE (2019)

Opportunity Assessment and Feasibility Study of IoT-Based Smart Farming in Bangladesh for Meeting Sustainable Development Goals

Nowshin Alam

Abstract The growing world population has placed increased pressure on the agricultural sector on a global scale, and all over the world, efforts are being made to increase food production. Smart farming is an Internet of Things (IoT)-based approach that optimizes productivity in terms of quality and quantity without compromising the farmers' economical circumstances or adding to their workloads. In this paper, the scope of smart farming has been considered in the perspective of Bangladesh where the Internet coverage is still not very reliable, and majority of field workforce are victims of poverty and illiteracy. Despite such barriers in the realization of smart farming, there have been several public and private projects aimed at gradually transforming the agricultural sector through the adaptation of sensor usage, IoT-based monitoring and satellite tracking. The feasibility of such endeavors has been reviewed in this paper, and the current implementation challenges have been discussed with some suggestions on possible solutions. The effect of such digital agriculture on the economy and the environment has also been linked to multiple Sustainable Development Goals (SDG) defined by the United Nations General Assembly. Toward the end, a conceptual framework for a low-cost smart farming system has been proposed that mainly comprises a number of ESP32 microcontrollers for collecting sensor data, a Raspberry Pi for hosting the database of sensor readings and the web application for viewing them from any device connected to the same network as the Pi. The wireless communication is performed over Wi-Fi, LoRa and GSM protocols.

Keywords Internet of Things · IoT · Smart agriculture · Smart farming

N. Alam (✉)
American International University-Bangladesh, 408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
e-mail: nowshin.alam@aiub.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_52

1 Introduction

In the last century, the world population has grown to a staggering number of more than 7 billion which has been forecasted to reach 9.6 billion by 2050 [1]. The whole world is facing the challenge of sustaining an ever-growing population with a limited quantity of natural reserves, out of which food is the most critical resource. To circumvent the looming threat of food shortage, all the countries worldwide are currently focusing their efforts into improving the efficiency and productivity of the agricultural sector, but due to geographical properties (such as temperature, landscape quality, weather, occurrence of natural disasters, etc.), not all countries are equally suitable for growing food. Furthermore, unstable factors such as government policies, political environment and ratio of urbanized versus rural areas impact the availability of cultivatable land. In the current situation where the traditional practices in agriculture are becoming obsolete over time, the emerging field of smart farming and digitalized agriculture may contribute to the solution for the current problems and future challenges regarding food production on a macroscopic scale.

The two integral factors of the smart farming approach are the usage of sensors and software-based monitoring [2]. The application of sensors and intelligent software improves quality and quantity of production of crop, livestock and fisheries, and the use of big data and machine learning aids post-harvest management and market growth. The authors of [3] expected the IoT technology used in smart farms to be capable of bringing a revolution in the agriculture industry and save the sector from its fate of being the weakest link in the economic value chain of many countries. Instead of how traditional farming causes the farmers to spend 70% [3] of their time monitoring the crops instead of doing actual field work requiring human wisdom, smart farming techniques enable farmers to properly monitor the state of farmlands and collect all necessary data accurately and automatically with minimal effort. This leads to a productivity increase as the farmers can focus their energy where it is needed and make informed decisions about crop selection, irrigation, disease and pest management, safety measures against natural disasters, etc. Modern agriculture methods contribute in the journey toward sustainable agriculture in this manner and justify the integration of smart farming into the production pipeline. However, the use of Internet and interconnected devices in smart farming rises new concerns, such as uncertainty in data security and availability [4], lack of skilled labor capable of operating new technology and sparseness of funding [5]. Not all methods of smart farming are equally applicable to each country due to the vast difference in climate conditions, technological advancement and financial prosperity. Even though agricultural sector is directly linked to the livelihood of over half of the population in Bangladesh, smart farming is still a very new concept in this country that still requires significant improvement in the farming infrastructure, training and education of the low-income rural population, expansion of high-speed Internet connectivity and governmental investment policies.

This paper tries to identify the feasible ways technology can contribute to the agricultural sector of Bangladesh in regards to sustainable farm management. For this purpose, the scope of smart farming in the modern world is discussed in Sect. 2, and the current state of smart farming in Bangladesh is summarized in Sect. 3. Section 4 discusses the major challenges hindering the immediate and widespread implementation of smart farming technology in Bangladesh, and Sect. 5 proposes a conceptual framework of a low-cost smart farming system prototype providing maximum utility from Bangladesh perspective while keeping the system affordable for small- and medium-scale farmers. Finally, Sect. 7 concludes this work by summarizing the takeaway from this research and discussing future action plans.

2 Smart Farming Scope

Smart farming is the approach of using modern technologies like IoT, sensors, robotics, UAVs and artificial intelligence to optimize agricultural output while utilizing minimal resources and human labor. The process can be considered a holistic approach that amalgamates information and environmental science as well as computer and software engineering and is further supported by technologies such Geographical Information System (GIS) and Global Positioning System (GPS) [6]. The goal of smart farming is to ensure high productivity, to create informed action plans that are adaptive and efficient and to mitigate negative effects to the environment such as groundwater depletion, greenhouse gas (GHG) emission and water pollution.

Out of the several aspects of smart farming, precision agriculture is a core component that utilizes localized, specific data to precisely locate crop issues and helps the farmer make informed decisions to increase crop yields and reduce risks. GPS and GIS can be of practical use in providing guidance in the handling of agricultural equipment such as tractors, sprayers and harvesters, and variable-rate technology (VRT) can control items like seeders and spreaders. Additionally, proper use of ICT tools built upon Internet of Things (IoT) is crucial to successfully implement smart farming as it facilitates the autonomous collection and analysis of agricultural data. Also in more advanced smart farming solutions, machine learning algorithms are utilized to identify patterns in data and predict important information about the crops and the farming environment [2].

2.1 Relationship Between Smart Farming and SDG Goals

The 2030 Agenda for Sustainable Development by United Nations General Assembly identified seventeen Sustainable Development Goals (SDGs) to promote education and growth across communities all over the world. Rahman et al. [5] state that, agricultural production needs to be doubled per acre by 2030 in order to achieve the SDGs related to agriculture, and smart farming can contribute to fulfilling this goal.

SDG 2. To combat hunger as per the "No hunger" target of SDG 2, the production capacity of the agriculture sector must be increased to ensure it can feed the population. As land expansion not possible in many densely populated countries including Bangladesh, increasing the resource utilization efficiency through smart farming is a more realistic approach.

SDG 12. Additionally, smart farming would prevent the overuse of resources by the implementation of smart irrigation systems and location-wise fertilizer applications. This would contribute to the "responsible consumption and production of resources" demanded by SDG 12.

SDG 11. "Sustainable communities" mentioned in SDG 11 would require a sustainable agricultural infrastructure that is prepared against natural disasters and mitigates the negative effect of human activities on the environment, both of which are aided by the IoT- and AI-based predictive measures of smart farming.

SDG 7. Lastly, integration of solar and wind power into the smart farms would contribute to ensuring "affordable and clean energy" requirement of SDG 7.

2.2 Technologies Employed in Smart Farming

Smart precision farming involves focusing of effort to correct application areas by letting the farmer make decisions for an individual area or object, with the operation scope being per square meter or even per plant or animal rather than for a whole field. The precise measurement of plant health can prevent unnecessary usage of water, fertilizer and pesticide. Similarly, individual animals can receive the specialized care they need, leading to early disease detection and prevention of outbreaks. As no single source can provide adequate information to a farming system for making informed decisions, a multitude of technologies is utilized to collect and process the data [6] (Fig. 1 and Table 1).

2.3 Smart Farming in World Perspective

From studying the history of countries that have already fully embraced the smart farming scheme, a massive increase in productivity can be observed. Compared to how 11 million farmers satisfied the food demand of 76 million Americans in the year of 1900, only 6.5 million farmers are fulfilling the food demand of 321 million humans in 2017 [2]. In 2018, the smart farming industry had a global market estimated at 7.5 billion USD and was dominated by United States and the Netherlands [4], but now South Korea has made rapid advance in the field through the application of cutting-edge technologies in vertical indoor farms and cultivating environments at innovative locations such as the unused spaces at subway stations in Seoul [7]. Michels et al. [8] report the frequent use of drones in Europe as a non-invasive method for gathering detailed data, but find limited application in pesticide and

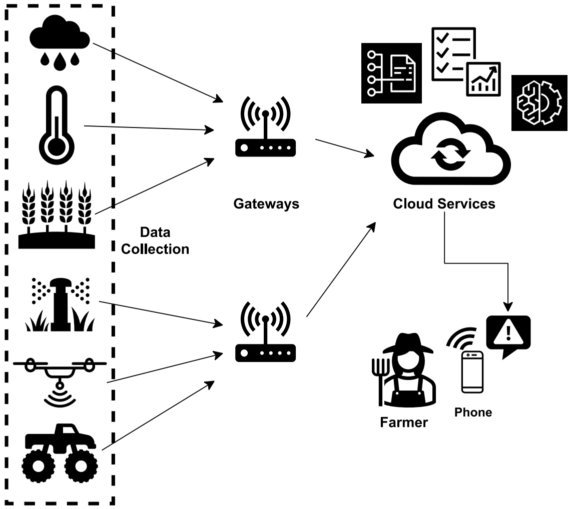


Fig. 1 Smart farming process

Table 1 Components of smart farming

Sensors	Temperature, soil moisture, water level, sunlight, humidity, pH
Software	Specialized software solutions for certain purposes such as frontend application to see data collected from farm
Connectivity	Wi-Fi, Cellular, Bluetooth, ZigBee, LoRa
Location	GPS, Satellite
Robotics	Autonomous tractors, agricultural drones
Algorithm	AI and machine learning algorithms

fertilizer spraying due to legal constraints. The Ministry of New and Renewable Energy (MNRE) has introduced the PM-KUSUM Farmers' Scheme in India, where farmers and other agricultural communities are provided subsidies for installing solar pumps. This would achieve the combined effect of contributing to grid as well as promoting green energy as no more fossil fuels will be used.

3 Smart Farming in Bangladesh

The irrigation technologies used before the independence of Bangladesh employed methods that were either inefficient or potentially harmful to the environment in long term, like swing buckets, drones and deep tube wells. Eventually, the agricultural sector grew more mechanized with the introduction of shallow and deep tube wells running on diesel engines and tractors for tilling. Newer techniques like Alternate Wetting and Drying (AWD) is being adopted by farmers to minimize irrigation water consumption without reducing the harvest [5]. However, simple mechanization of agriculture alone cannot achieve the level of growth the country is aspiring for.

3.1 Recent IoT Applications for Smart Farming

While there are still many challenges that need to be resolved before smart farming becomes widespread in Bangladesh, several public and private sector projects have already dipped their toes in the smart farming industry. Multiple dairy and livestock farms in the country have embraced the use of bolus sensors that report the health condition of cattle at all times. Dutch Dairy Ltd at Shatghariya, Munshiganj, is a large-scale cow farm operating since 2018 that employs modern cattle rearing methods to maximize productivity and utilizes smart technologies such as automatic temperature control through fan and sprinklers in addition to the constant cattle condition monitoring through sensors [9]. Digi Cow and Shurjomukhi Pranisheba, in collaboration with Grameenphone, have aided farms with the setup of such IoT services [5]. Shurjomukhi Pranisheba also introduced IoT and RFID-enabled cattle insurance.

Recently, agricultural fintech company iFarmer has also began to provide Fitbit cattle sensor and soil moisture sensor to their clients [10]. They launched the Sofol app in 2020 that helps with the operation of their Sofol agriculture finance and supply chain platform. Once NID verified, small farmers sign up in the app through a facilitator, their finances are logged in the iFarmer's database which allows the iFarmer supply chain process to connect them with investors on the basis of a credit scoring model, as well as provide them IoT-related support like weather updates and disaster alerts.

3.2 Recent Software and Mobile Applications for Smart Farming

In addition to the development of physical farming methodologies, there has been quite a number of phone and web applications that provide assistance in agriculture efforts. The Department of Agricultural Extension (DAE) provides the Agvisely web platform and the BAMIS portal mobile app for weather and flood forecasting as well as farming advice. CZ-IIS, a GIS-based web application by Bangladesh Agricultural Research Council (BARC), provides a detailed guide to farmers to help them choose which crop is suitable for their lands and allows access of the same functions through a mobile app called Khamari. However, it should be noted that some of the websites and apps designed by the government departments are works in progress that do not function properly yet and are not user-friendly enough for farmers to use.

Additionally, Bangladesh Meteorological Department (BMD) also provides weather-related information and real-time observations of parameters such as rainfall, carbon dioxide, soil moisture and water quality for many locations at their website. Such information would be very helpful in smart farming if provided in the form of Application Programming Interface (API) for IoT application developers, like the National Weather Service (NWS) API provided by USA government.

3.3 Possible Smart Farming Application Sectors

The fish farming sector in Bangladesh has started to adopt technologies like Recirculated Aquaculture System (RAS) and Biofloc [5] to reduce waste and farm fish with low mortality rate. Entrepreneurs have begun to combine fish farming with vertical and soilless (aquaponic) farms implanted in rooftops [11] which is further boosting productivity. The environmentally friendly floating farms on rectangular rafts on water by farmers of Gopalganj, Barisal and Pirojpur have attracted attention worldwide [12]. The UN's Food and Agricultural Organization recognized these floating gardens to be a global heritage in December 2015, and the government of Bangladesh also recognized the potential of implementing these water farms in other parts of the country to allow farming during periods of waterlogging and flooding each year. The 1.6 million USD project that took place in 50 locations around the country generated consistent yields, but more research and more rural market development is required to successfully duplicate this farming method at locations with different climates and farming infrastructure. All of the above farming approaches, especially the floating farms that are difficult to physically navigate, may greatly benefit from IoT applications and remote sensing technology.

4 Current Challenges and Possible Solutions

4.1 Internet Connectivity

The telecommunication sector has undergone a big revolution in the recent years, but the network coverage in rural areas is still not satisfactory, as evidenced by World Bank 2019 report's claim of only 13% of total population of Bangladesh having Internet access [13]. As smart farming largely relies on Internet of things and many IoT functions require a network connection to provide their full functionality, the proper deployment of smart farming solutions is hindered by unreliable network conditions [7]. However, on August 2021, BTRC reported 115.41 million mobile Internet subscribers and 10.05 million wired Internet (ISP and PSTN) users based on ISP data reports and market analysis, whereas the total population of the country is 166,817,904 as of October 2021 based on Worldometer elaboration of the most recent United Nations (UN) data. While the rate of growth of Internet coverage seems promising, passing large volumes of data such as camera feeds may be difficult due to the expensive Internet data packages. To promote smart farming practices, the government should provide subsidy for farmers and improve the telecommunication infrastructure to provide uninterrupted Internet connectivity to even rural locations.

4.2 Lack of Funding

Many of the farming methods employed in first world countries are too expensive for Bangladesh. The cheapest agricultural drone starts from 1, 500 USD (approximately 1,30,000 BDT) which is beyond the buying capacity of small- or medium-scale farmers. While government has started providing subsidy and loans to farmers for setting up solar powered irrigation systems in their fields, the poorest farmers who can barely make ends meet are not eligible to apply. Wealthy farmers have an advantage over them in regards to funds, so government needs to ensure that the smart farming promotion campaign does not overly rely on the farmers' personal funds or encourage discrimination. Farmer's cooperatives can play a role in helping farmers share the cost of solar pumps through communal use. An average 18.5 kW solar water pump is capable of lifting 25–30 lac liters of water per day and cover 130 bighas (32.6 ha) of land, which is enough for our country where the average farm size has decreased to 0.6 ha [10]. Leasing smart farms to farmers on short-term basis like Korea are also a feasible method, and expensive equipment like drones can be utilized by sharing devices among local farming communities.

4.3 Lack of Education and Training

Majority of the farmers in Bangladesh are impoverished with minimal education and are incapable of utilizing technology for agricultural development. The size of skilled workforce is quite small and educated young people are uninterested in agriculture-related professions due to the financial uncertainty that is associated with these jobs [2]. In case of hardware malfunction, rural farms can suffer severe damage in the absence of skilled technical experts. Systematic training through agricultural extension service is vital for the implementation of smart farming. The potential of smart farming should be publicized through television and social media, so more young and educated people are attracted by the promises of per person labor decrease and demand for ICT skills in the sector.

4.4 Farmland Fragmentation

In a country like Bangladesh with rich crop diversity, farmers own multiple small lands of uneven arrangement which are hard to design IoT solutions around. Additionally, the farmers' tendency to consider whole fields as single farming units makes IoT application even harder. Rahman et al. [5] suggest assigning farmlands into smaller management zones based upon their soil quality, but they also observe that there exists little literature providing guidance on the process of segmenting the areas.

4.5 Climate Instability

The climate of Bangladesh varies largely across different regions. The northern region suffers from droughts and thus sports water deficiency and high temperatures, whereas the southern region, especially the haor areas (wetland ecosystem that is a bowl or saucer shaped shallow depression), encounters frequent floods and experiences saline soil and a lack of quality water due to water contamination. This greatly increases the necessity to perform intensive surveys and planning before any IoT system design and no universal plan can be applied.

However, once a predictive AI-based smart farming system is set into place, it would be able to warn the farmers against upcoming floods or droughts based on climate data trends, so farmers would be able to take necessary precautions like harvesting available crops or switching to a different crop variety. The use of solar energy would also contribute saving water level and reducing fossil fuel consumption.

4.6 Government Policies

There have been many isolated projects by both public and private sectors that involved smart farming [7], but there is still very little public interest toward the topic. Government initiatives and public–private partnership are required for the scale of widespread realization of smart farming. A bigger percentage of GDP should be invested in the agricultural sector, and the barrier to smart farming should be lowered through incentives and policies such as adjustments to exchange, agricultural product prices, tariffs and subsidies [5]. The Civil Aviation Authority of Bangladesh (CAAB) also imposes many restrictions on the usage of drones in Bangladesh where any drone above 5 kg requires permission at least 45 days in advance, making agricultural use of UAVs difficult.

5 Proposed Framework

In this section, a smart farming system is proposed that collects data from nodes placed at different corners of a small farming field, issues warnings about weather and soil condition and controls the irrigation system. The observed parameters at each node are soil moisture, temperature, humidity, air pressure and water level. Soil moisture is observed at every location, whereas the remaining properties are observed at a single node as these readings does not vary significantly in different locations. The collected data from the nodes is stored using a Raspberry Pi 2 micro-controller connected to the Internet. As Internet connectivity is unreliable in many rural areas, the sensor information is accessible through both a web application and phone messages (Fig. 2).

5.1 *Justification for Device Choice*

The justification for using two different types of devices is the fact that the Raspberry Pi's General Purpose Input/Output (GPIO) interface does not provide analog inputs. Using only Raspberry Pi would require the installation of an expensive HAT extension board [14] that supports adding sensors or a MCP3008 analog-to-digital converter (ADC) chip to convert the analog sensor inputs to digital. To keep the cost to a minimum while keeping the design simple, we have used the powerful but affordable ESP32 microcontroller which hosts 18 analog-to-digital converters on board. However, out of these pins, only 8 pins belonging to ADC1 channel can be used as analog pins collecting sensor readings as the ADC2 pins are unavailable for use during Wi-Fi operation. ESP32 also allows easy configuration of a mesh network of

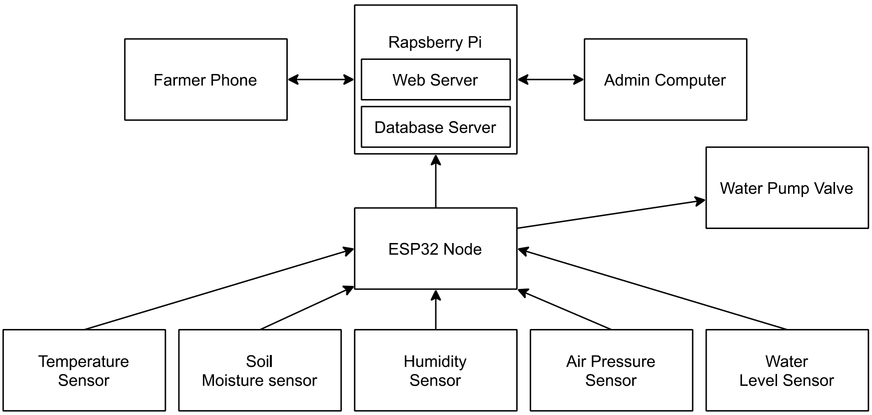


Fig. 2 Block diagram (showing the node connected to all sensors) of the proposed smart farming system

the nodes, allowing the extension of the wireless connectivity to distances normally not covered by a router.

5.2 Operation

Two servers are set up in the Raspberry Pi. A web server runs a GUI frontend application for viewing the sensor data from any device connected to the same network as the Raspberry Pi. It is built using Python's Flask micro-web framework. The Pi also hosts a database server that locally stores the sensor data over time in a MySQL database, unlike previous approaches where the data was only saved in cloud storage [15] and data collection was suspended in case of power outage. In the case of no Internet connectivity, the Raspberry Pi will keep recording the information in its database and send the farmer periodic messages using the GSM Board (SIM800) (Fig. 3).

Furthermore, the Raspberry Pi will study the sensor values and alert the farmer in both the web app and through phone message if abnormalities are found in the data. For example, sudden drop in temperature and air pressure in addition to humidity increase would result in a warning about the possibility of rain, storm or cyclone. The soil moisture and the water level readings will also be utilized to turn the solenoid valves in the farm's drip irrigation system on or off.

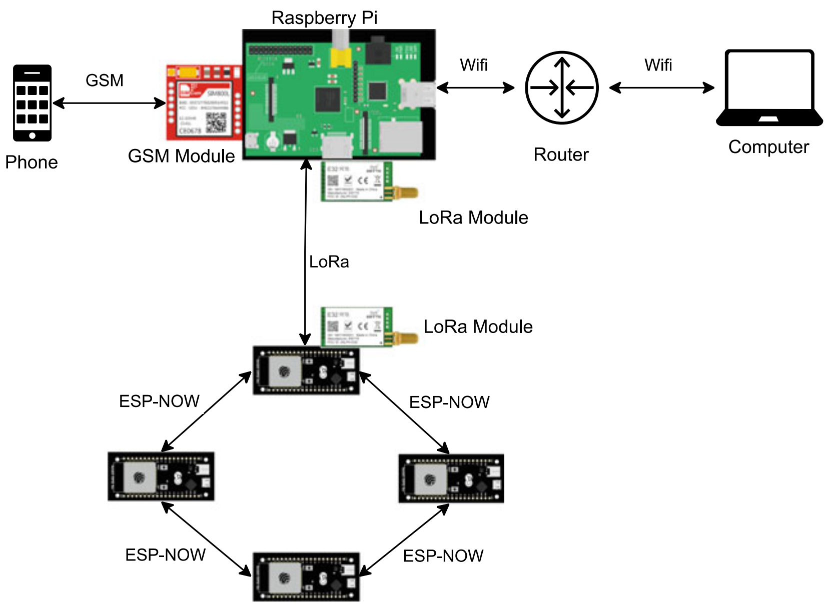


Fig. 3 Mesh connection of multiple ESP32 nodes placed over a field and wireless communication interfaces in different parts of the system

Table 2 Cost analysis

Component name	Local price (BDT)
Raspberry Pi 2 Model B	2450
SIM800 GSM Module	450
ESP32 NodeMCU Board × 4	790 × 4
SX1278 LoRa Module × 2	990 × 2
Temperature + Humidity Sensor × 1	180 × 1
Water Level Sensor × 1	500 × 1
BMP280 Pressure Sensor × 1	290 × 1
Capacitive Moisture Sensor × 4	299 × 4

The mesh connected ESP32 microcontrollers will communicate using the versatile ESP-NOW [16] protocol that is very fast in exchanging small messages like sensor readings, whereas the ESP32 that will communicate with the Raspberry Pi (assumed to be far away from the field, presumably at the building where the administrator of smart farm is) will use Long Range (LoRa) transceivers similar to the deployment process in [17]. The choice is due to LoRa's suitability for low-power applications that run on battery and impressive range that extends up to kilometers. The total price of the hardware (excluding the cost of wiring and enclosures) is estimated to be only BDT 10,206 which is within the budget of even small farmers. The cost can be further reduced for smaller fields that require less sensors and do not need LoRa.

Currently, the software development using Python (for Raspberry Pi setup and building the web page), C (for ESP32 setup) and SQL (for setting up the database) is in progress for proposed framework above. After the development of the initial prototype, flood and drought prediction capabilities will be improved by using a machine learning algorithm. The machine learning model will be deployed using Tensorflow on the smart farm administrator's computer connected to the same network as the Raspberry Pi (Table 2).

6 Conclusion

Though the process has been gradual, smart farming and digital agriculture are paving the way for the next Green Revolution by embracing the integration of ICT approaches such as precision farming, IoT, robotics and AI. In Bangladesh, a variety of stakeholders other than the farmers (for example, agricultural product suppliers, policy makers, researchers, etc.) are taking part in the development of innovative applications related to precision agriculture. It is of utmost importance that they are not hindered by digital divide or poor farming management. Thanks to governmental policies' increasing support of entrepreneurship projects in the last couple of years [18], entrepreneurs and educated farmers are much more driven and hopeful about approaching smart farming. Considering the potential of smart farming for ensuring a productive and sustainable society and achieving the SDGs in the country, government should place increased focus on improvement of the agricultural infrastructure for the sake of successful implementation of future smart farming-related programs.

References

1. Schönfeld, M.V., Heil, R., Bittner, L.: Big data on a farm—smart farming. In: Big Data in Context. pp. 109–120 (2018). https://doi.org/10.1007/978-3-319-62461-7_12
2. Smart Farming: The future of agriculture. https://agssbd.org/smart-farmingthe-future-of-agriculture/ (2021)
3. Ayaz, M., Ammad-Uddin, M., Sharif, Z., Mansour, A., Aggoune, E.-H.M.: Internet-of-things (IoT)-based smart agriculture: toward making the fields talk. IEEE Access. **7**, 129551–129583 (2019). https://doi.org/10.1109/ACCESS.2019.2932609
4. Gupta, M., Abdelsalam, M., Khorsandroo, S., Mittal, S.: Security and privacy in smart farming: challenges and opportunities. IEEE Access. **8**, 34564–34584 (2020). https://doi.org/10.1109/ACCESS.2020.2975142
5. Rahman, A., Ali, R., Kabir, S.N.: Agricultural mechanization in Bangladesh: status and challenges towards achieving the sustainable development goals (SDGs). AMA Agric. Mech. Asia Afr. Lat. Am. **51**, 106–120 (2020)
6. Terence, S., Purushothaman, G.: Systematic review of internet of things in smart farming. Trans. Emerg. Telecommun. Technol. **31**, (2020). https://doi.org/10.1002/ett.3958
7. Seraj, S.: Smart age shaping agriculture and food security. https://www.thedailystar.net/supplements/29th-anniversary-supplements/governance-development-and-sustainable-bangladesh/news/smart-age-shaping-agriculture-and-food-security (2020)
8. Michels, M., von Hobe, C.-F., Weller von Ahlefeld, P.J., Musshoff, O.: The adoption of drones in German agriculture: a structural equation model. Precis. Agric. **22**, 1728–1748 (2021). https://doi.org/10.1007/s11119-021-09809-8
9. Seraj, S.: Hi-tech dairy farm in Munshiganj shows bright prospect. https://www.thedailystar.net/country/news/hi-tech-dairy-farm-munshiganj-shows-bright-prospect-1864027 (2020)
10. Thorsteinsdóttir, H., Bell, J.M., Bandyopadhyay, N.: Cultivating small and medium-sized firms: entrepreneurship development, gender, and technology in Bangladesh, Cambodia, Ethiopia and Senegal. United Nations Technology Bank for the Least Developed Countries, Canada (2021)
11. Azad, K.N., Salam, M.A., Azad, K.N.: Aquaponics in Bangladesh: current status and future prospects. J. Biosci. Agric. Res. **7**, 669–677 (2016). https://doi.org/10.18801/jbar.070216.79
12. Sunder, K.: The remarkable floating gardens of Bangladesh. https://www.bbc.com/future/article/20200910-the-remarkable-floating-gardens-of-bangladesh
13. The World Bank. https://data.worldbank.org/indicator/IT.NET.USER.ZS
14. Navulur, S., S.C.S., Sastry, A.N., Giri Prasad, M.: Agricultural management through wireless sensors and internet of things. Int. J. Electr. Comput. Eng. IJECE. **7**, 3492 (2017). https://doi.org/10.11591/ijece.v7i6.pp3492-3499
15. Chieochan, O., Saokaew, A., Boonchieng, E.: Internet of things (IOT) for smart solar energy: A case study of the smart farm at Maejo University. In: 2017 International Conference on Control, Automation and Information Sciences (ICCAIS). pp. 262–267. IEEE, Chiang Mai (2017)
16. Choong, J.H.: Disease detection using sensors with raspberry Pi through wireless sensor network in vegetable farms. http://eprints.utar.edu.my/3062/1/fyp_CN_2018_CJH_-_1605821.pdf (2018)

17. Islam, N., Ray, B., Pasandideh, F.: IoT based smart farming: are the LPWAN technologies suitable for remote communication? In: 2020 IEEE international conference on smart internet of things (SmartIoT), pp. 270–276. IEEE, Beijing, China (2020). https://doi.org/10.1109/SmartIoT49966.2020.00048
18. Deb, U.: Agricultural transformation in Bangladesh: Extent, drivers and implications. In: 15th National Conference of the Bangladesh Agricultural Economists Association (BAEA) on "Transformation of Agricultural Sector in Bangladesh: 21st Century, p. 36. , Dhaka, Bangladesh (2016)

An Explainable Alzheimer's Disease Prediction Using EfficientNet-B7 Convolutional Neural Network Architecture

Sobhana Jahan and M. Shamim Kaiser

Abstract Alzheimer's disease is a neurocognitive disease that results from the brain shrinking and brain tissue dying over time. It gradually erodes memory, thinking skills, and the ability to carry out the most basic tasks. The use of an MRI to evaluate brain atrophy is thought to be a reliable way to diagnose and track the progression of Alzheimer's disease. In such studies, deep learning architecture provides outstanding results. One drawback of deep learning is that it necessitates a large number of datasets to train the model. Another drawback is the black-box nature, and due to this nature, doctors, patients, and the general public are doubtful of the model's prediction results. To remove these problems, this study proposes a novel Gradient-weighted class activation mapping (Grad-CAM)-based explanation of Alzheimer's disease prediction, using the EfficientNet-B7 convolutional neural network architecture. Here, the data augmentation technique is used to make the small dataset suitable for the model. In this work, we have performed a classification between different stages of Alzheimer's disease which are Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Alzheimer's disease (AD). Using the base dataset and augmented dataset, the achieved accuracy for four-class classification is 73.47% and 91.76%, respectively. The resulting accuracy for AD versus CN, CN versus EMCI, CN versus LMCI, AD versus EMCI, AD versus LMCI, and EMCI versus LMCI using augmented dataset is 97.72%, 97.70%, 98.01%, 95.12%, 96.70%, and 96.40%, respectively.

Keywords Deep learning · Transfer learning · Explainable AI · Data augmentation · Gradient-weighted class activation mapping

---

S. Jahan (✉)
Information and Communication Technology, Bangladesh University of Professionals, Dhaka, 1216, Bangladesh
e-mail: sobhanajahanlisa@gmail.com

M. Shamim Kaiser
Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, 1342, Bangladesh
e-mail: msKaiser@juniv.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9_53

1 Introduction

Alzheimer's disease (AD) is a neurological disorder in which memory loss and cognitive decline are caused due to the death of brain cells. AD and other dementias are thought to affect at least 50 million people worldwide. As Asia is the world's most populous continent, a report estimates that the number of dementia patients in this region will reach nearly 71 million by 2050 [1]. In Western societies, AD is the most common cause of dementia, and around 5.5 million people in the USA are affected by it [2]. There is currently no cure for Alzheimer's disease, but there are medications available to slow the progression of the disease.

An early Alzheimer's diagnosis increases the chances of being benefitted from treatment. It also creates the opportunity for future care and treatment. For early diagnosis of AD, it is necessary to identify those patients who are at Mild Cognitive Impairment (MCI) stage. Researchers discovered that people with MCI are more likely than those without it to develop AD or related dementia. Schmidtke et al. [3] found that among all MCI patients, 44% had converted to AD after a mean delay of 19 months.

For confirming an Alzheimer's diagnosis or ruling out some other potential causes, medical professionals also use brain scans such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), or Positron Emission Tomography (PET). As MRI is one of the most powerful neuroimaging techniques used in deep learning-based AD prediction models, in this work we have used MRI data collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI). From the ADNI cohort, MRI images of a total of four types of participants are collected and those types are Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and AD.

Since 2013, deep learning has gotten a lot of attention in Alzheimer's disease detection research. When compared to general machine learning techniques, deep models are more accurate in detecting Alzheimer's disease [4]. Transfer learning is a previously trained deep learning model for representing features of a new dataset. A pre-trained model is typically trained on a large dataset like ImageNet, and the weights acquired from the trained model can be used with one custom neural network for other applications. So, it is commonly used to save time and resources. EfficientNet is a convolutional neural network that uses a compound coefficient to evenly scale all width, resolution, depth, or dimensions.

Here in this research work, the EfficientNet-B7 model is used for training and preprocessing the input data. As the EfficientNet-B7 is a Convolutional Neural Network (CNN) model, it requires a huge amount of data for providing a good result. This work performed data augmentation to remove the problem of data size. In data analysis, data augmentation refers to techniques for increasing the amount of data by trying to add slightly altered copies of existing data or creating new synthetic data from existing data.

Due to the black-box nature of deep learning models, general people are unable to know the reasons behind the model's prediction. That is why doctors and patients do not believe in the deep learning models' prediction. To overcome this issue, this work has proposed the gradient-weighted class activation mapping (Grad-CAM) technique to visualize the decision-making parts of the MRI image.

The main contributions of this work are as follows:

1. We have proposed a data augmentation technique for increasing the size of the training and testing dataset.
2. We have presented the EfficientNet-B7 model to classify the stages of AD.
3. We have used Grad-CAM to convert the black-box decisions into explainable ones so that anyone can understand the reasons for any prediction.

The rest of the paper is structured as follows: Section 2 includes related works, Sect. 3 contains the proposed methodology, Sect. 4 comes up with performance comparisons, and Sect. 5 includes the concluding remarks.

2 Related Works

Nanni et al. [5] presented a comparison study between ensemble SVM and ensemble transfer learning methods for predicting Alzheimer's in the early stage using MRI. It was observed that the ensemble SVM model provides better AUC than ensemble transfer learning for AD versus CN and CN versus MCIc prediction, which is 93.2% and 90.6%, respectively. For MCIc versus MCInc prediction, ensemble transfer learning was better performing than machine learning and that value is 70.6%. By using a multi-slice and multi-model ensemble approach, Kang et al. [6] proposed an AD prediction model with the help of CNNs. When categorizing MCI versus AD, CN versus MCI, and CN versus AD, this ensemble method obtained accuracy values of 77.19%, 72.36%, and 90.36%, respectively.

Wee et al. [7] used a spectral graph-CNN with cortical thickness and geometry to detect MCI and AD. The graph-CNN had shown accuracy rate of AD versus CN at 85.8%, AD versus EMCI at 79.2%, CN versus LMCI 69.3%, EMCI versus LMCI 60.9%, LMCI versus AD 65.2%, and EMCI versus CN 51.8% for the ADNI-2 dataset. Mehmood et al. [8] separated EMCI, CN, LMCI, and AD using the VGG architecture. The best accuracy was 98.73% for AD versus CN. For LMCI versus EMCI patients, the evaluated accuracy was 83.72%.

Using MRI data, Ji et al. [9] proposed an early diagnosis of Alzheimer's disease using CNN. In that paper, three base ConvNets were used, and these three classifiers are eNASNet, eResNet-50, and eMobileNet. Using the ADNI dataset, the eNASNet model had achieved a sensitivity of 80.56%, a specificity of 94.00%, an AUC of 0.96%, and an accuracy of 88.37% for CN versus MCI estimation. Naz et al. [10] used the AlexNet, VGG, Inceptionv3, GoogLeNet, ResNet, DenseNet, and InceptionResNet architectures to detect AD. Using the ADNI dataset, VGG-16 achieved 97.06% and 98.89% accuracy for MCI versus CN and AD versus CN, respectively. VGG-19 achieved the highest accuracy of 99.27% for MCI versus AD.

3 Proposed Methodology

3.1 Dataset Acquisition

The Alzheimer Disease Neuroimaging Initiative-ADNI dataset is the gold standard for Alzheimer disease detection [11]. ADNI was established in 2003 under the direction of Dr. Michael W. Weiner and is funded through public–private cooperation. The ADNI is a multisite, longitudinal study aimed at developing clinical, genetic, imaging, and biomarkers for the early detection of AD. ADNI, ADNI GO, ADNI 2, and ADNI 3 in these three databases, there are 1800 subjects, both females and males [12].

In this study, a subset of the ADNI dataset is used to train, validate, and test different architectures of CNN classifiers. Axial T2-Star weighted longitudinal MRI data of 85 AD, 85 CN, 76 EMCI, and 76 LMCI is chosen for this work. As a result, our base dataset contains a total of 1633 MRI data points.

3.2 Data Pre-processing and Data Augmentation

For data pre-processing, we have performed rotation and resizing operations. After that we have performed image augmentation which is an artificial technique for increasing the size of a training dataset by constructing modified versions of the images. Horizontal flip, 5 units rotation, 0.1 width shift, 0.1 height shift, and 0.1 zooms are used in this work to perform augmentation. After the data augmentation, we have achieved a large dataset of 6392 data including 1420 CN, 1562 LMCI, 1740 EMCI, and 1645 AD. Here, 85% image data are used as training data, and the rest are used as testing data. The algorithm of the data augmentation technique is mentioned below:

Input: Original image data θ;
Output: Augmented training image set λ;
Define $\theta = \lambda$;
For each image sample I^k (fi) $in\theta$ do
  For offset $\beta_\zeta = -\upsilon : \upsilon_0 : \upsilon$ do
    Add β_ζ to the true azimuth $\hat{\beta}$, $\beta' = \hat{\beta} + \beta_\zeta$;
    Generate the beamformed signals;
    Generate the feature ν_ζ^k;
    $\lambda = \lambda$ U ν_ζ^k;
  End
End

Here, the original data is denoted as θ, which is disturbed at the time of the feature extraction to get the augmented features λ. Apparently, for each sample in θ, we have found the augmented feature ν_ζ^k (here, is the sample index in θ) by proposing an offset angle β_ζ to the real azimuthangle $\hat{\beta}$. The augmented signal is calculated by the azimuth angle $\beta' = \hat{\beta} + \beta_\zeta$.

3.3 Classification Model

EfficientNet is one of the most effective CNN models that achieve state-of-the-art accuracy on both ImageNet and common image classification transfer learning tasks, which was first created by Tan and Le in 2019 [13]. EfficientNet has eight variants (B0 to B7), and the width, depth, and resolution of each variant are handpicked and demonstrated to produce better results. Considering the performance of all eight variants, it is found that EfficientNet-B7 is the best performing. So, this work is constructed using EfficientNet-B7. To simplify the model, we divided the total 813 layers into six modules and one stem. In this case, stem includes the following actions or layers: Input, Rescaling, Normalization, Zero Padding, Conv2D, Batch Normalization, and Activation. Module 1's layers are as follows: Depthwise Conv2D, Batch Normalization, and Activation. Similarly, the other five modules are constructed, as shown in Fig. 1.

Figure 2 comes up with the complete architecture of the EfficientNet-B7 model. As it's name suggests, there are total seven blocks in this model. The layers of modules and stems are stated in Fig. 1. In the first block, module 1 takes input from the stem. The output of module 1 goes to the module 3. Module 4's input is the output of module 3. Module 4's output goes to the module 1. The output of module 1 goes to the module 3, and this module 3's output further goes to the module 5. Add layer takes the output of modules 5. Module 1, 3, 5, and Add stays inside a loop, and this loop continues two times. After the execution of this loop, the final result from block 1 transfers to block 2's module 2. Figure 2 explains all the modules and their sequence for creating the prediction model. We have constructed the last three layers manually by inserting dense layer. Among those three layers, the first two layers are defined as Dense layers with the unit of '128' and the activation function is 'ReLu.' For four-class classification, the output layer or the last layer contains the unit value '4' along with the 'softmax' activation function. For two-class classification, the last layer contains the unit value '2' along with the 'sigmoid' activation function. For remaining layers, we have used the EfficientNet-B7 model's default layers.

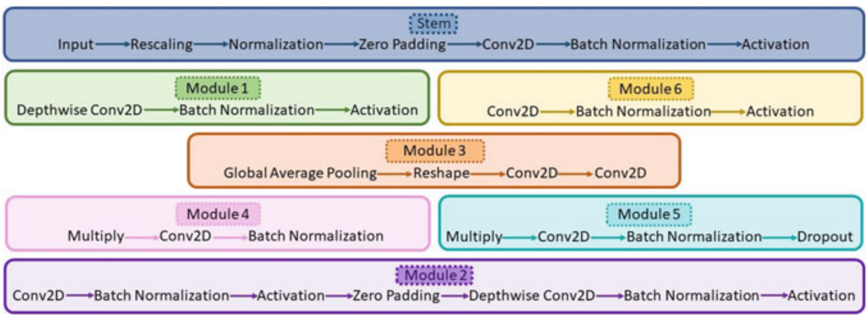


Fig. 1 Modules and stem of EfficientNet-B7 architecture

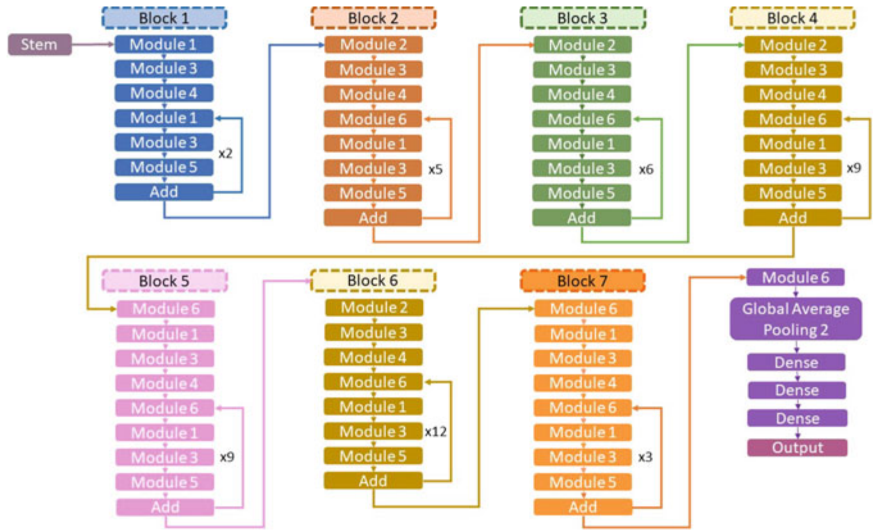


Fig. 2 Complete architecture of EfficientNet-B7 model

3.4 *Explain Ability of the Model*

Grad-CAM [14] is a technique for improving the transparency of CNN-based models by visualizing the input territories that are 'crucial' for the model's predicted results [15]. Grad-CAM generates a class-specific heatmap based on an input image [16]. Through that heat map, one can easily understand the most dominating (red-colored) and less dominating (blue-colored) portions for any decision.

The main operation of Grad-CAM is described in the following sections. To begin, the oth rectified feature map Po can be denoted as

$$P^o = \frac{1}{Z} \sum_r \sum_s B^o_{r,s} \tag{1}$$

Here, Z is the feature map's number of pixels, r and s are the feature map's row and column indexes, and $B^o_{r,s}$ is the score of the pixel in the rth row and sth column. The class C score (R^C) also recognized as the class C's score before the softmax layer.

$$R^C = \sum_o \beta^C_o P^o \tag{2}$$

Here, β^C_o is the oth feature map's weight. The first activation map ($W^C_{\text{Grad-CAM}}$) produced by Grad-CAM is

$$W^{C}_{\text{Grad-CAM}} = \text{RELU}\left(\sum_{o} \beta^{C}_{o} P^{o}\right) \tag{3}$$

The gradient method can be used to analyze the weights of a pixel. The procedure can be represented as follows:

$$\beta^{C}_{o} = \frac{\delta R^{C}}{\delta P^{o}} = \frac{\delta R^{C}}{\delta B^{o}_{r,s}} \frac{\delta B^{o}_{r,s}}{\delta P^{o}} = \frac{\delta R^{C}}{\delta B^{o}_{r,s}}.Z \tag{4}$$

The average of weights of every pixel in the feature map is the weight of the oth feature map. The *o*th feature map's weight is

$$\frac{1}{Z}\sum_{r}\sum_{s}\beta^{C}_{o} = \frac{1}{Z}\sum_{r}\sum_{s}\frac{\delta R^{C}}{\delta B^{o}_{r,s}}.Z \tag{5}$$

Equation 5 can be simplified because the weights of every pixel in the *o*th feature map has to be the same.

$$\frac{1}{Z}.Z.\beta^{C}_{o} = \frac{1}{Z}\sum_{r}\sum_{s}\frac{\delta R^{C}}{\delta B^{o}_{r,s}}.Z = \sum_{r}\sum_{s}\frac{\delta R^{C}}{\delta B^{o}_{r,s}} \tag{6}$$

Using Grad-CAM, the normalized localization map (attention map or heat map) can be defined as

$$S = \frac{1}{Z}\sum_{r}\sum_{s}\sum_{o}\beta^{C}_{o} B^{o}_{r,s} \tag{7}$$

3.5 Proposed Explainable Alzheimer's Disease Prediction Using EfficientNet-B7 Architecture

Figure 3 presents the model architecture of this research work. Here, T2 star MRI images of four types (CN, AD, EMCI, and LMCI) of participants are used. Initially for this task, 1633 MRI images are used. For preparing the dataset, we have performed image rotation, resizing, and data augmentation. To make sufficient dataset for the EfficientNet-B7 model, we have performed augmentation. After augmentation, the number of data has become 6392 which is overall satisfactory for the EfficientNet-B7 CNN model. Then, we have performed the classification operation using the EfficientNet-B7 pre-trained CNN model and found the classification result. For visualizing the decision-making portions of the MRI images, we have used the Grad-CAM algorithm. The last convolutional layer of the EfficientNet-B7 named '*topconv*' is passed to the Grad-CAM algorithm and the final output is shown using the heat map.

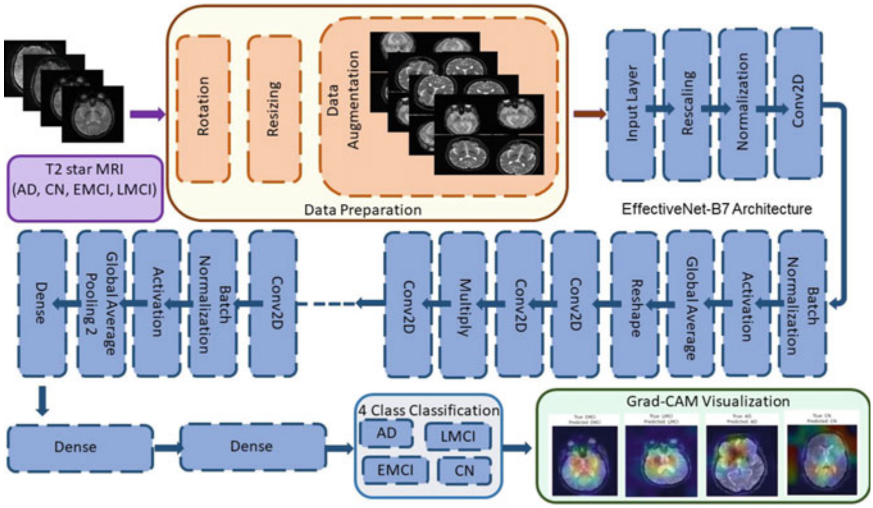


Fig. 3 Proposed model architecture for Alzheimer's prediction and visualization

4 Performance Analysis

4.1 Performance Metrics

A confusion matrix is an overview of classification problem's prediction results. From confusion matrix, we can calculate the Accuracy, Precision, Recall, and F1-score using True positive (TP), True negative (TN), False Positive (FP), and False Negative (FN) values. The rules of Accuracy, Precision, Recall, and F1-score are written below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

4.2 Data Augmentation Versus Without Augmentation

For analyzing the effect of the augmented dataset, we have performed four class classifications on the augmented dataset and the base dataset. Without augmentation, the overall accuracy value is 73%, and with the augmented dataset, this accuracy is

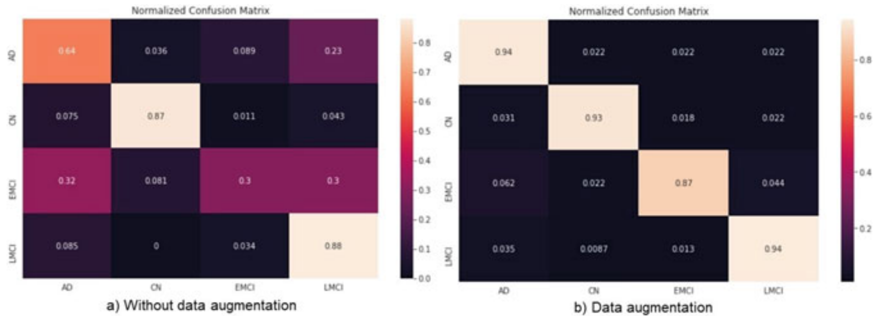


Fig. 4 Confusion matrix of the model using augmented dataset and non-augmented dataset

Table 1 Accuracy comparison among various transfer learning models for AD versus CN versus EMCI versus LMCI classification

Classifier	Accuracy (%)	Classifier	Accuracy (%)	Classifier	Accuracy (%)
InceptionResNet-V2	74.45	MobileNet-V2	87.49	ResNet-34	83.73
ResNet-50	84.78	ResNet-101	90.71	ResNet-152	89.89
VGG-16	88.74	VGG-19	90.09	DenseNet-121	84.05
DenseNet-201	88.63	Xception	75.91	NASNetMobile	70.59
Inception-V3	67.67	EfficientNet-B0	88.63	EfficientNet-B1	91.35
EfficientNet-B2	91.66	EfficientNet-B3	90.51	EfficientNet-B4	90.30
EfficientNet-B5	88.11	EfficientNet-B6	90.82	EfficientNet-B7	**91.76**

91.76%. Figure 4 comes up with the normalized confusion matrix where it is clear that the performance using the augmented dataset is far better than the base dataset. Using the augmented dataset the highest Precision value achieved for EMCI and that is 95%. We found the highest Recall value of 94% for AD and LMCI class.

4.3 Comparison Between Transfer Learning Models

There are several outstanding existing transfer learning models which are usually used in AD versus CN versus EMCI versus LMCI classification. Table 1 shows that the accuracy of EfficientNet-B7 is higher than other models.

4.4 Accuracy Comparison with Other Recent Works

From Table 2, we have found that for all six types of binary classification our proposed model achieved highest accuracy scores than others. For CN versus LMCI

Table 2 Accuracy comparison among similar works on six binary classifications

Binary classification	Korolev et al. [17] (ResNet) (%)	Shakeri et al. [18] (multi layer perceptron) (%)	Our proposed work (data augmentation + EfficientNet-B7 +Grad-CAM) (%)
AD versus CN	80	84	**97.72**
CN versus EMCI	56	56	**97.70**
CN versus LMCI	61	59	**98.01**
AD versus EMCI	63	81	**95.12**
AD versus LMCI	59	67	**96.70**
EMCI versus LMCI	52	63	**96.40**

Table 3 Accuracy comparison among various recently published similar works for AD versus CN classification

Authors	Model	AD versus CN (%)
Helaly et al. [19]	2D-M2IC	97.11
Wee et al. [20]	graph-CNN	81.0
Khvostikov et al. [21]	3D inception	93.3
Dan et al. [17]	CNN-EL	84
Zhang et al. [18]	TRRA+EfficientNet-B1	93
Our proposed Work	Augmentation+ EfficientNet-B7+Grad-CAM	**97.72**

classification, we have found 98.01% accuracy, which is better than other binary classifications. Here, all the authors have used the same dataset collected from the ADNI repository.

Table 3 consists the comparison between various recently published work using the AD and CN MRI data collected from the ADNI repository. From this table, it is clear that the accuracy of our proposed work is higher than others and that value is 97.72%.

4.5 Result Visualization

As we have already stated that, we have used Grad-CAM technique to visualize the reason behind any specific classification result. Figure 5 shows the Grad-CAM visualization of this model. The heatmap represents the level of dominance or attention for any classification. Here, the red-colored areas are highly involved for the decision making.

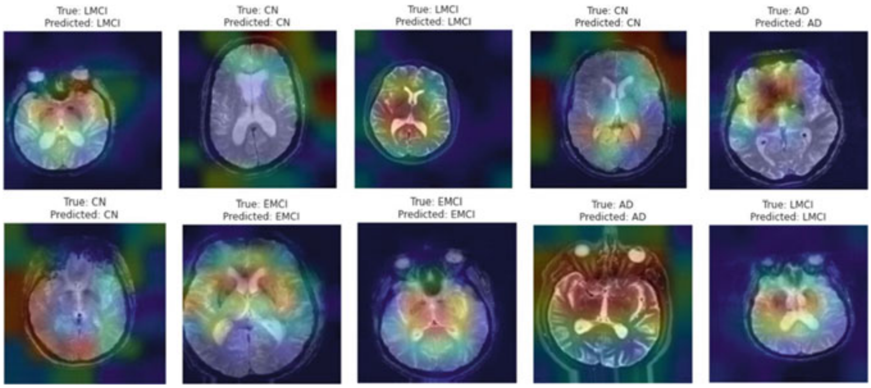


Fig. 5 Explaination of prediction result using Grad-CAM

5 Conclusion

In this study, we have proposed a Grad-CAM-based EfficientNet-B7 model for Alzheimer's prediction, where the data augmentation technique is used to increase the number of MRI data as required for deep learning. With the help of this approach, we can easily provide predictions on CN, AD, EMCI, and LMCI stages. Our proposed approach removes the black-box nature of the deep learning-based prediction model and also outperforms many recently published works in this field. For AD versus CN versus EMCI versus LMCI classification, our proposed approach achieved 91.76% accuracy. The accuracies achieved from the augmented dataset for AD versus CN, CN versus EMCI, CN versus LMCI, AD versus EMCI, AD versus LMCI, and EMCI versus LMCI are 97.72%, 97.70%, 98.01%, 95.12%, 96.70%, and 96.40%, respectively. In the future, we will try to implement a deep learning model for early-stage AD prediction. For improving the model's performance we are likely to implement an ensemble CNN model.

Acknowledgements This work was supported by the research funds of the Information and Communication Technology (ICT) division of the Government of the People's Republic of Bangladesh in 2020–2021.

References

1. Dementia in the Asia Pacific Region, pp. 64. [Online]. Available: https://www.alzint.org/u/Dementia-Asia-Pacific-2014.pdf. Accessed 05 Nov 2021
2. Mayeux, R., Stern, Y.: Epidemiology of alzheimer disease. Cold Spring Harbor Perspect. Med. **2**(8), a006239 (2012)
3. Schmidtke, K., Hermeneit, S.: High rate of conversion to Alzheimer's disease in a cohort of amnestic MCI patients. Int. Psychogeriatrics **20**(1), 96–108 (2008)

4. Ebrahimighahnavieh, M.R., Luo, S., Chiong, R.: Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. Comput. Meth. Prog. Biomed. **187**, 105242 (2020)
5. Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., Castiglioni, I.: Comparison of transfer learning and conventional machine learning applied to structural brain MRI for the early diagnosis and prognosis of Alzheimer's disease. Front. Neurol. **11**, 576194 (2020)
6. Kang, W., Lin, L., Zhang, B., Shen, X., Shuicai, W.: Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis. Comput. Biol. Med. **136**, 104678 (2021)
7. Wee, C.Y., Liu, C., Lee, A., Poh, J.S., Ji, H., Qiu, A.: Cortical graph neural network for AD and MCI diagnosis and transfer learning across populations. NeuroImage: Clinic. **23**, 101929 (2019)
8. Mehmood, A., Yang, S., Feng, Z., Wang, M., Ahmad, A.S., Khan, R., Maqsood, M., Yaqub, M.: A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. Neuroscience **460**, 43–52 (2021)
9. Ji, H., Liu, Z., Yan, W.Q., Klette, R.: Early diagnosis of Alzheimer's disease using deep learning. In: Proceedings of the 2nd International Conference on Control and Computer Vision, ICCCV 2019, pp. 87–91. Association for Computing Machinery, New York, NY, USA (2019)
10. Naz, S., Ashraf, A., Zaib, A.: Transfer learning using freeze features for Alzheimer neurological disorder detection using ADNI dataset. Multimedia Syst. **28**(1), 85–94 (2022)
11. Weber, C., Carrillo, M., Jagust, W., Jack, C., Shaw, L., Trojanowski, J., Saykin, A., Beckett, L., Sur, C., Rao, N., Mendez, P., Black, S., Li, K., Iwatsubo, T., Chang, C.-C., Sosa, A., Rowe, C., Morris, J., Weiner, M.: The worldwide Alzheimer's disease neuroimaging initiative: ADNI-3 updates and global perspectives. In: Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 7 (2021)
12. Clifford, R., Jack, Jr., Matt, A., Bernstein, Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, D., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magnet. Resonance Imag. **27**(4), 685–691 (2008)
13. Tan, M., Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs, stat]. arXiv: 1905.11946 (2020)
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128**(2), 336–359 (2020). arXiv: 1610.02391
15. Vinogradova, K., Dibrov, A., Myers, G.: Towards interpretable semantic segmentation via gradient-weighted class activation mapping. arXiv:2002.11434 [cs, eess]. arXiv: 2002.11434 (2020)
16. Grad-CAM: Gradient-weighted Class Activation Mapping
17. Frontiers | Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning | Neuroscience
18. Zhang, F., Pan, B., Shao, P., Liu, P., Shen, S., Yao, P., Xu, R.X.: An explainable two-dimensional single model deep learning approach for Alzheimer's disease diagnosis and brain atrophy localization. arXiv:2107.13200 [cs, eess]. arXiv: 2107.13200 (2021)
19. Helaly, H.A., Badawy, M., Haikal, A.Y.: Deep learning approach for early detection of Alzheimer's disease. Cognitive Comput. (2021)
20. Cortical graph neural network for AD and MCI diagnosis and transfer learning across populations—ScienceDirect
21. Khvostikov, A., Aderghal, K., Krylov, A., Catheline, G., Benois-Pineau, J.: 3D Inception-based CNN with sMRI and MD-DTI data fusion for Alzheimer's Disease diagnostics (2018)

Author Index

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

Md. S. Hossain et al. (eds.), *The Fourth Industrial Revolution and Beyond*, Lecture Notes in Electrical Engineering 980, https://doi.org/10.1007/978-981-19-8032-9