

Vishal Goar

Manoj Kuri

Rajesh Kumar

Tomonobu Senju *Editors*

# Advances in Information Communication Technology and Computing

Proceedings of AICTC 2019

# **Lecture Notes in Networks and Systems**

**Volume 135**

## **Series Editor**

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,  
Warsaw, Poland

## **Advisory Editors**

Fernando Gomide, Department of Computer Engineering and Automation—DCA,  
School of Electrical and Computer Engineering—FEEC, University of Campinas—  
UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering,  
Bogazici University, Istanbul, Turkey

Derong Liu, Department of Electrical and Computer Engineering, University  
of Illinois at Chicago, Chicago, USA; Institute of Automation, Chinese Academy  
of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering,  
University of Alberta, Alberta, Canada; Systems Research Institute,  
Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering,  
KIOS Research Center for Intelligent Systems and Networks, University of Cyprus,  
Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong,  
Kowloon, Hong Kong

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

**\*\* Indexing: The books of this series are submitted to ISI Proceedings, SCOPUS, Google Scholar and Springerlink \*\***

More information about this series at <http://www.springer.com/series/15179>

Vishal Goar · Manoj Kuri ·  
Rajesh Kumar · Tomonobu Senju  
Editors

# Advances in Information Communication Technology and Computing

Proceedings of AICTC 2019



Springer

*Editors*

Vishal Goar  
Department of Computer Application  
Government Engineering College  
Bikaner, Rajasthan, India

Rajesh Kumar  
Department of Electrical Engineering  
Malaviya National Institute of Technology  
Jaipur, Rajasthan, India

Manoj Kuri  
Department of Electronics  
and Communication  
Government Engineering College  
Bikaner, Rajasthan, India

Tomonobu Senju  
Department of Electrical  
and Electronics Engineering  
University of the Ryukyus  
Nishihara, Japan

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-981-15-5420-9

ISBN 978-981-15-5421-6 (eBook)

<https://doi.org/10.1007/978-981-15-5421-6>

© Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are reserved 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

# **Committees**

## **Chief Patron**

Mr. Hemant Kumar Bohra, Industrialist, Bohra Group of Companies

## **Chairman**

Dr. Jaiprakash Bhamu, Principal, Government Engineering College Bikaner

## **General Chair**

Dr. Vishal Goar, Assistant Professor, Government Engineering College Bikaner

## **Conference Chair**

Mr. Manoj Kuri, Assistant Professor, Government Engineering College Bikaner

## **Program Chair**

Dr. S. K. Bishnoi, Associate Professor, Government Engineering College Bikaner

## **Coordinator**

Mr. Ajay Choudhary, Assistant Professor, Government Engineering College Bikaner

Mr. Pawan Tanwar, Assistant Professor, Government Engineering College Bikaner

## **Co-coordinator**

Dr. Abdul Jabbar Khilji, Assistant Professor, Government Engineering College Bikaner

Mr. Rajendra Singh Sekhawat, Assistant Professor, Government Engineering College Bikaner

## **Proceeding Chair**

Dr. O. P. Jakhar, Associate Professor, Government Engineering College Bikaner

Dr. Nilanjan Dey, Techno India College of Technology, Kolkata

Dr. Amit Joshi, Vice Chairman, ACM Udaipur Professional Chapter

Dr. Ankur Dumeja, Graphic Era (Deemed to be University), Dehradun

**Publicity Chair**

Dr. Munesh Trivedi, NIT Agartala  
Dr. Brojo Kishore Mishra, GIET University, Gunupur, India  
Mr. Abdul Sammad, Marudhar Engineering College Bikaner

**Apex Chair**

Dr. Anand Sharma, Secretary, CSI Lakshmangarh Chapter  
Dr. Chiranjit Lal Chowdhary, VIT, Vellore  
Dr. Vishal Goyal, Punjabi University, Patiala  
Dr. L. C. Bishnoi, Government Polytechnic College Bikaner  
Dr. Niranjanamurthy M., M. S. Ramaiah Institute of Technology, Bengaluru  
Dr. S. Swamynathan, College of Engineering Guindy, Anna University, Chennai  
Mr. Prasant Joshi, Department of Computer Science, Government Polytechnical College, Bikaner, Rajasthan

**Track Chair**

Dr. Radha Mathur, Government Engineering College Bikaner  
Dr. Devendra Gahlot, Government Engineering College Bikaner  
Dr. Rakesh Poonia, Government Engineering College Bikaner  
Dr. Ankur Goswami, Government Engineering College Bikaner  
Dr. Sanjay Tejasvee, Government Engineering College Bikaner  
Mr. Ganesh Singh, Government Engineering College Bikaner  
Mr. Rituraj Soni, Government Engineering College Bikaner  
Mr. Ranu Lal Chouhan, Government Engineering College Bikaner  
Mr. Kunal Bhusan Ranga, Government Engineering College Bikaner  
Mr. Surya Prakash Takhar, Government Engineering College Bikaner

**Web Designing Chair**

Mr. Kapil Vyas, Government Engineering College Bikaner  
Dr. Rajpal Choudhary, Acharya Shree Nanesh Samta Mahavidyalaya, Danta, Chittorgarh

**Advisory Committee**

Prof. Valentina Emilia Balas, "Aurel Vlaicu" University of Arad, Romania  
Prof. Sanjay Misra, Covenant University, OTA, Nigeria  
Dr. Hussain Falih Mahdi, Computer and Software Engineering Department, College of Engineering, University of Diyala, Iraq  
Dr. Vinaye Armoogum, University of Technology, Mauritius  
Dr. Haruna Chiroma, University of Malaya  
Dr. Pierre Clarel Catherine, University of Technology, Mauritius, La Tour Koenig, Pointe-aux-Sables  
Prof. Aboul Ella Hassanien, Cairo University, Egypt  
Prof. Manuel Castro, National University of Distance Education, Spain  
Prof. Olga Poleshchuk, Bauman State Technical University, Mytichi Branch, Moscow, Russian Federation  
Prof. Ho Chin Kuan, Faculty of Information Technology, Multimedia University, Malaysia

Prof. Shawulu Nggada, Higher Colleges of Technology, UAE  
Dr. David Alan Grier, The George Washington University, Washington, Columbia  
Dr. Arpan K. Kar, IIT Delhi  
Dr. Arun Raj Kumar P., NIT, Puducherry  
Prof. Paresh V. Virparia, Department of Computer Science, Sardar Patel University, Vallabh Vidyanagar  
Dr. Parameshachari B. D., Professor and Head, GSSSIETW, Mysuru  
Dr. Dilip Kumar Yadav, National Institute of Technology, Jamshedpur, India  
Dr. Seema S., M. S. Ramaiah Institute of Technology, Bengaluru, India  
Dr. R. Sunder, MET's School of Engineering, Thrissur, Kerala, India  
Dr. Jibi Abraham, College of Engineering, Pune  
Dr. Mayank Aggarwal, GKV University, Haridwar  
Prof. Himanshu Aggarwal, Punjabi University, Patiala  
Prof. Yanxia Sun, University of Johannesburg, Johannesburg  
Dr. Vishal Kumar, Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand

### **Program Committee**

Prof. Ting-Peng Liang, National Chengchi University, Taipei, Taiwan  
Nedia Smairi, CNAM Laboratory, France  
Prof. Subhadip Basu, Visiting Scientist, The University of Iowa, Iowa City, USA  
Prof. Abrar A. Qureshi, Ph.D., University of Virginia, USA  
Prof. Louis M. Rose, Department of Computer Science, University of New York, USA  
Prof. Dr. Ricardo M. Checchi, University of Massachusetts, MA, USA  
Prof. Brent Waters, University of Texas, Austin, TX, USA  
Prof. Prasun Sinha, Ohio State University, Columbus, OH, USA  
Prof. N. M. van Straalen, VU University Amsterdam, Amsterdam, The Netherlands  
Prof. Rashid Ansari, University of Illinois, USA  
Prof. Russell Beale, School of Computer Science—Advanced Interaction, University of Birmingham, England  
Prof. Dan Boneh, Computer Science Department, Stanford University, CA, USA  
Prof. Alexander Christea, University of Warwick, London, UK  
Prof. Mustafizur Rahman, Endeavor Research Fellow, Australia  
Prof. Hoang Pham, Rutgers University, Piscataway, NJ, USA  
Prof. Ernest Chulantha Kulasekere, University of Moratuwa, Sri Lanka  
Prof. Shashidhar Ram Joshi, Institute of Engineering, Pulchowk, Nepal  
Dr. Ashish Rastogi, Higher College of Technology, Muscat, Oman  
Dr. Aynur Unal, Standford University, USA  
Prof. Ahmad Al-Khasawneh, The Hashemite University, Jordan  
Dr. Bharat Singh Deora, JRN RV University, India  
Prof. Jean Michel Bruel, Departement Informatique IUT de Blagnac, Blagnac, France  
Prof. Ngai-Man Cheung, Assistant Professor, University of Technology and Design, Singapore

- Prof. J. Andrew Clark, Computer Science University of York, UK  
Prof. Babita Gupta, College of Business California State University, CA, USA  
Prof. Shuiqing Huang, Department of Information Management, Nanjing Agricultural University, Nanjing, China  
Prof. Yun-Bae Kim, SungKyunKwan University, South Korea  
Prof. Sami Mnasri, IRIT Laboratory, Toulouse, France  
Prof. Anand Paul, The School of Computer Science and Engineering, South Korea  
Dr. Krishnamachar Prasad, Department of Electrical and Electronic Engineering, Auckland, New Zealand  
Dr. Haibo Tian, School of Information Science and Technology, Guangzhou, Guangdong, China  
Er. Kalpana Jain, CTAE, Udaipur, India  
Prof. Philip Yang, Pricewaterhouse Coopers, Beijing, China  
Prof. Sunarto Kampus, UNY, Yogyakarta, Indonesia  
Dr. Ashok Jetawat, CSI Udaipur Chapter, India  
Dr. Neetesh Purohit, Member SIG-WNs CSI, IIT Allahabad, India  
Mr. Jeril Kuriakose, Manipal University, Jaipur, India  
Prof. R. K. Bayal, Rajasthan Technical University, Kota, Rajasthan, India  
Prof. Martin Everett, University of Manchester, England  
Prof. Feng Jiang, Harbin Institute of Technology, China  
Dr. Savita Gandhi, Professor, Gujarat University, Ahmedabad, India  
Mr. Chintan Bhatt, Changa University, Gujarat, India  
Prof. Feng Tian, Virginia Polytechnic Institute and State University, USA  
Prof. XiuYing Tian, Instrument Laboratory, Yangtze Delta Region Institute of Tsinghua University, Jiaxing, China  
Prof. Xiaoyi Yu, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China  
Prof. Abdul Rajak A. R., Department of Electronics and Communication Engineering, Birla Institute of Technology and Sciences, Abu Dhabi  
Mr. Ajay Choudhary, IIT Roorkee, India  
Dr. Manju Mandot, CSI Udaipur Chapter, India  
Prof. D. A. Parikh, Head, CE, LDCE, Ahmedabad, India  
Dr. Paras Kothari, Samarth Group of Institutions, Gujarat, India  
Dr. Harshal Arolkar, Immd. Past Chairman, CSI Ahmedabad Chapter, India  
Mr. Bhavesh Joshi, Advent College, Udaipur, India  
Prof. K. C. Roy, Principal, Kautaliya, Jaipur, India  
Dr. Mukesh Shrimali, Pacific University, Udaipur, India  
Dr. Sanjay M. Shah, GEC, Gandhinagar, India  
Dr. Sonam Mishra, M-SIG-WNs, CSI, KEC Dwarahat, Uttarakhand, India  
Dr. Chirag S. Thaker, GEC, Bhavnagar, Gujarat, India  
Mr. Nisarg Pathak, SSC, CSI, Gujarat, India  
Mrs. Meenakshi Tripathi, MNIT, Jaipur, India  
Prof. S. N. Tazi, Government Engineering College, Ajmer, Rajasthan, India  
Shuhong Gao, Mathematical Sciences, Clemson University, Clemson, South Carolina

- Sanjam Garg, University of California, Los Angeles, CA  
Faiez Gargouri, Sfax University, Tunisia, North Africa  
Dr. A. Garrett, Department of Mathematics, Computing, and Information Sciences, Jacksonville State University, Jacksonville, Alabama  
Leszek Antoni Gasieniec, University of Liverpool, Liverpool, England  
Ning Ge, School of Information Science and Technology, Tsinghua University, Beijing, China  
Garani Georgia, University of North London, UK  
Hazhir Ghasemnezhad, Electronics and Communication Engineering Department, Shiraz University of Technology, Shiraz, Iran  
Andrea Goldsmith, Professor of Electrical Engineering, Stanford University, CA  
Dr. Saeed Golmohammadi, Assistant Professor in University of Tabriz, Tabriz, Iran  
Prof. K. Gong, School of Management, Chongqing Jiaotong University, Chongqing, China  
Crina Gosnan, Associate Professor Department of Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania  
Mohamed Gouda, The University of Texas, Computer Science Department, Austin, TX  
Mihai Grigore, Department for Management, Technology and Economics Group for Management, Information Systems, Zurich, Switzerland  
Cheng Guang, Southeast University, Nanjing, China  
Venkat N. Gudivada, Weisburg Division of Engineering and Computer Science, Marshall University Huntington, Huntington, West Virginia  
Prof. Wang Guojun, School of Information Science and Engineering of Zhong Nan University, China  
Prof. Nguyen Ha, Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatchewan, Canada  
Dr. Z. J. Haas, School of Electrical Engineering, Cornell University, Ithaca, New York  
Prof. Mohand Said Hacid, Lyon University, France  
Prof. Haffaf Hafid, University of Oran, Oran, Algeria  
Prof. M. Tarafdar Hagh, Department of Electrical Engineering, Islamic Azad University, Ahar, Iran  
Ridha Hamdi, University of Sfax, Tunisia, North Africa  
Prof. Dae Man Han, Green Home Energy Center, Kongju National University, Republic of Korea  
Prof. Xiangjian He, University of Technology, Sydney, Australia  
Prof. Richard Heeks, University of Manchester, Manchester, UK  
Mr. Walid Khaled Hidouci, Ecole nationale Supérieure d'Informatique, Algeria  
Prof. Achim Hoffmann, School of Computer Science and Engineering, The University of New South Wales, Australia  
Ma Hong, Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China  
Hyehyun Hong, Department of Advertising and Public Relations, Chung-Ang University, South Korea

- Qinghua Hu, Harbin Institute of Technology, China  
Honggang Hu, School of Information Science and Technology, University of Science and Technology of China, P.R. China  
Fengjun Hu, Zhejiang Shuren University, Zhejiang, China  
Dr. Qinghua Huang, School of Electronic and Information Engineering, South China University of Technology, China  
Chiang Hung-Lung, China Medical University, Taichung, Taiwan  
Kyeong Hur, Department of Computer Education, Gyeongin National University of Education, Incheon, Korea  
Wen-Jyi Hwang, Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei  
Gabriel Sebastian Ioan Ilie, Computer Science and Engineering Department, University of Connecticut, Mansfield, Connecticut  
Sudath Indrasinghe, School of Computing and Mathematical Sciences, Liverpool John Moores University, Liverpool, England  
Ushio Inoue, Department of Information and Communication Engineering, Engineering Tokyo Denki University, Tokyo, Japan  
Dr. Stephen Intille, Associate Professor College of Computer and Information Science and Department of Health Sciences, Northeastern University, Boston, MA  
Dr. M. T. Islam, Institute of Space Science, Universiti Kebangsaan Malaysia, Selangor, Malaysia  
Lillykutty Jacob, Professor, Department of Electronics and Communication Engineering, NIT, Calicut, Kerala, India  
Anil K. Jain, Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan  
Dagmar Janacova, Tomas Bata University in Zlin, Faculty of Applied Informatics, Nam. T. G., Czech Republic, Europe  
Kairat Jaroenrat, Faculty of Engineering at Kamphaeng Saen, Kasetsart University, Bangkok, Thailand  
Don Jyh-Fu Jeng, Assistant Professor, Institute of International Management, National Cheng Kung University, Taiwan  
Minseok Jeon, Department of Computer Science, Yonsei University, Seoul, South Korea  
Prof. Guangrong Ji, College of Information Science and Engineering, Ocean University of China, Qingdao, China  
Yoon Ji-Hyeun, Department of Computer Science, Yonsei University, Seoul, South Korea  
Zhiping Jia, Computer Science and Technology, Shandong University, Jinan, China  
Liangxiao Jiang, Department of Computer Science, China University of Geosciences, Wuhan, China  
David B. Johnson, Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania  
Prof. Chen Junning, Electronic Information and Engineering, Anhui University, Hefei, China

- Seok Kang, Associate Professor, University of Texas, San Antonio, TX  
Ghader Karimian, Assistant Professor, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran  
S. Karthikeyan, Department of Information Technology, College of Applied Science, Sohar, Oman, Middle East  
Michael Kasper, Fraunhofer Institute for Secure Information Technology, Germany  
L. Kasprzyczak, Institute of Innovative Technologies EMAG, Katowice, Poland  
Zahid Khan, School of Engineering and Electronics, The University of Edinburgh, Mayfield Road, Scotland  
Jin-Woo Kim, Department of Electronics and Electrical Engineering, Korea University, Seoul, Korea  
Muzafer Khan, Computer Sciences Department, COMSATS University, Pakistan  
Jamal Akhtar Khan, Department of Computer Science College of Computer Engineering and Sciences, Salman bin Abdulaziz University, Kingdom of Saudi Arabia  
Kholaddi Kheir Eddine, University of Constantine, Algeria  
Dr. Fouad Khelifi, School of Computing, Engineering and Information Sciences, Northumbria University, Newcastle upon Tyne, England  
Shubhalaxmi Kher, Arkansas State University, College of Engineering, Jonesboro, Arkansas  
Sally Kift, James Cook University, Townsville, Queensland  
Sunkyun Kim, Department of Computer Science, Yonsei University, Seoul, Korea  
Leonard Kleinrock, Computer Science Department, University of California, Los Angeles, CA  
Dirk Koch, School of Computer Science, University of Manchester, Manchester, England  
Zbigniew Kotulski, Warsaw University of Technology, Faculty of Electronics and Information Technology Institute of Telecommunications, Warszawa, Poland  
Ray Kresman, Bowling Green State University, Bowling Green, OH, USA  
Ajay Kshemkalyani, Department of Computer Science, University of Illinois, Chicago, IL  
Madhu Kumar, Associate Professor, Computer Engineering Department, Nanyang Technological University, Singapore  
Anup Kumar, Professor, Director MINDS Laboratory, University of Louisville, KY, USA  
James Tin-Yau Kwok, Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong  
Zhiling Lan, Department of Computer Science, Illinois Institute of Technology, Chicago, IL  
Hayden Kwok-Hay So, Department of Electrical and Electronics Engineering, University of Hong Kong, Hong Kong  
K. G. Langendoen, Delft University of Technology, The Netherlands, Europe  
Michele Lanza, REVEAL Research Group, Faculty of Informatics, University of Lugano, Switzerland

- Shalini Batra, Computer Science and Engineering Department, Thapar University, Patiala, Punjab, India
- Shajulin Benedict, Director, HPCCloud Research Laboratory, St. Xavier's Catholic College of Engineering, Chunkankadai District, Nagercoil, Tamil Nadu
- Rajendra Kumar Bharti, Assistant Professor, Kumaon Engineering College, Dwarahat, Uttarakhand, India
- Prof. Murali Bhaskaran, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India
- Prof. Komal Bhatia, YMCA University, Faridabad, Haryana, India
- Prof. S. R. Biradar, Department of Information Science and Engineering, SDM College of Engineering and Technology, Dharwad, Karnataka
- A. K. Chaturvedi, Department of Electrical Engineering, IIT Kanpur, India
- Jitender Kumar Chhabra, NIT, Kurukshetra, Haryana, India
- Pradeep Chouksey, Principal, TIT College, Bhopal, Madhya Pradesh, India
- Chhaya Dalela, Associate Professor, JSSATE, Noida, Uttar Pradesh
- Jayanti Dansana, KIIT University, Bhubaneswar, Odisha
- Soura Dasgupta, Department of TCE, SRM University, Chennai, India
- Dr. Apurva A. Desai, Veer Narmad South Gujarat University, Surat, India
- V. Susheela Devi, Senior Scientific Officer, Department of Computer Science and Automation, Indian Institute of Science, Bengaluru
- Dr. Bikash Kumar Dey, Department of Electrical Engineering, IIT Bombay, Powai, Maharashtra
- Vijay Pal Dhaka, Jaipur National University, Jaipur, Rajasthan
- K. Bhattachary Dhruba, Department of Computer Science and Engineering, Tezpur University, Assam, India
- Mohammad Doja, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi
- Prof. Sagayaraj Francis, Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India
- K. Ganesh, Supply Chain Management—Center of Competence, McKinsey Knowledge Center India Private Limited, McKinsey & Company, Gurgaon, Haryana
- Dr. Vinit Grewal, Assistant Professor, Department of Electronics and Communication Engineering, Guru Nanak Dev University, Jalandhar, India
- P. S. Grover, Ex. Professor, Department of Computer Science, University of Delhi, New Delhi
- Prof. S. Hemalatha, Professor/CSE, PSNA College of Engineering and Technology, Dindigul, India
- Hazra Imran, Department of Computer Science, Jamia Hamdard, New Delhi, India
- S. Janakiraman, Professor Department of IT, Pondicherry University, Puducherry, India
- A. P. Kabilan, Principal, Vivekanandha College of Engineering and Technology, Tiruchengode, India
- V. Kavitha, Anna University, Chennai, India

- Rama Krishna, Department of Computer Science and Engineering, National Institute of Technical Teachers Training and Research, Chandigarh, India  
R. Krishnamoorthy, Dean, Anna University, Chennai, BIT Campus, Trichy, India  
Jagadeesh Kumar, Sri Krishna College of Technology, Coimbatore, India  
S. Maheswaran, Assistant Professor, Kongu Engineering College, India  
Dr. Sushil Kumar, School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India  
Amioy Kumar, Biometrics Research Laboratory, Department of Electrical Engineering, IIT Delhi, India  
S. Britto Ramesh Kumar, Department of Computer Science, St. Joseph's College, Tiruchirappalli, Tamil Nadu  
Manish Kumar, Assistant Professor, IIIT, Jhalwa, Allahabad, India  
T. Arun Kumar, Professor, School of Computing Science and Engineering, Software Systems Division, Vellore, Tamil Nadu, India  
Prof. Mohamed Ben Mohamed, University of Constantine, Algeria, North Africa  
Prof. Kristin P. Bennett, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, USA  
Prof. Abhir Bhalerao, Department of CSE, University of Warwick, London, UK  
Prof. Norman L. Biggs, London School of Economics, UK  
Prof. Liu Bing, Department of Computer Science, University of Illinois, Chicago, IL  
Prof. Qin Bo, Universitat Rovira i Virgili, Tarragona, Spain, Europe  
Prof. Fatima Boumahdi, Ouled Yaich Blida, Algeria, North Africa  
Prof. Nikolaos G. Bourbakis, Department of Computer Science and Engineering, Dayton, Ohio, Montgomery  
Prof. Mokrane Bouzghoub, Laboratoire PRISM, Versailles, France  
Prof. Alan Conrad Bovik, The University of Texas at Austin, Electrical and Computer Engineering, Austin, TX, USA  
Mr. Janez Brank, Institut Jozef Stefan, Ljubljana, Slovenia, Europe  
Prof. Torsten Braun, IAM, Neubrückstrasse, Bern, Switzerland  
Prof. Lionel Brunie, The Institut National des Sciences Appliquées de Lyon, Villeurbanne Cedex, France  
Prof. Dongbo Bu, Institute of Computing Technology, Chinese Academy of Science, Beijing, China  
Prof. Alister Burr, Electronics Department, University of York, York, North Yorkshire  
Luca Cagliero, Doc. Research Fellow, Department of Control and Computer Engineering, Turin, Italy  
Prof. Berrut Catherine, LIG Laboratory of Grenoble, Grenoble University, France  
Prof. K. P. Chan, Associate Professor, Department of Computer Science, University of Hong Kong  
Prof. Zhou Chao, The Civil Aviation Flight University of China, Chengdu, Sichuan, P. R. China  
Somayeh Mamizadeh Chatghayeh, Asia Business Clusters and Networks Development, Foundation Cooperation, Tehran, Iran

- Sandeep Chatterjee, CTO, Trace Systems, CA, USA  
Mr. Zhidao Chen, Chuangyuan M&E Co. Ltd., Changchun, China  
Prof. Yixing Chen, Shanghai Jiao Tong University, Shanghai, China  
Prof. Yawen Chen, University of Otago Clocktower, Dunedin, New Zealand  
Prof. Cailian Chen, Department of Automation, Shanghai Jiao Tong University, Shanghai, China  
Prof. Janson Cheng, University of Birmingham, England  
Prof. Chin-Wan Chung, Department of Electrical Engineering and Computer Science, Korea  
Prof. C. J. Chung, Department of Math and Computer Science, Lawrence Technological University, Southfield, Michigan  
Prof. Jonathan Clark, STRIDE Laboratory Mechanical Engineering, Tallahassee, Florida  
Prof. Thomas Cormen, Department of Computer Science Dartmouth College, Hanover, Germany  
Prof. Dennis D. Cox, Rice University, TX, USA  
Prof. Marcos Roberto da Silva Borges, Federal University of Rio de Janeiro, Brazil  
Prof. Gholamhossein Dastghaibyfard, College of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran  
Prof. Doreen De Leon, California State University, USA  
Bartel Van de Walle, University Tilburg, Tilburg, The Netherlands  
Prof. David Delahaye, Saint-Martin, Cedex, France  
Prof. Andrew G. Dempster, The University of New South Wales, Australia  
Prof. Alan Dennis, Kelley School of Business Indiana University, Bloomington, IN, USA  
Prof. Jitender Singh Deogun, Department of Computer Science and Engineering, University of Nebraska–Lincoln, NE, USA  
Dr. S. A. D. Dias, Department of Electronics and Telecommunication Engineering, University of Moratuwa, Sri Lanka  
Dr. Zhang Dinghai, Gansu Agricultural University, Lanzhou, China  
Prof. Ali Djebbari, Sidi Bel Abbes, Algeria  
Dr. P. D. D. Dominic, Department of Computer and Information Science, Universiti Teknologi Petronas, Tronoh, Perak, Malaysia  
Prof. David Douglas, Walton College of Business, University of Arkansas, USA  
Dr. Vishal Goyal, Punjabi University, Patiyala, Punjab, India

# Preface

The 2nd International Conference on *Advances in Information Communication Technology and Computing* (AICTC-2019) was held at Government Engineering College Bikaner, Rajasthan, India, during November 8–9, 2019. AICTC-2019 was organized by Government Engineering College Bikaner, Rajasthan, India, and supported by TEQIP-III. The main purpose of AICTC-2019 is to provide a leading edge, scholarly forum for researchers, engineers, and students alike to share their state-of-the-art research and developmental work in the broad areas of pervasive computing and communications. The conference will feature a diverse mixture of interactive forums and core technical sessions of high-quality cutting-edge research articles. The field of ICT and computing always deals with finding innovative solutions for problems by proposing different techniques, methods, and tools. The proceedings covers systems, paradigms, techniques, and technical reviews that employ knowledge and intelligence in a broad spectrum. **AICTC-2019 received around 262 submissions from 603 authors of eight different countries such as Taiwan, Sweden, Italy, Saudi Arabia, China, and many more.** Each submission has gone through the plagiarism check. On the basis of a plagiarism report, each submission was rigorously reviewed by at least two reviewers. Even some submissions have more than two reviews. On the basis of these reviews, 52 high-quality papers were selected for publication in this proceedings volume.

We are thankful to the speakers: Dr. Arpan K. Kar, IIT, Delhi, India, and Dr. Nilanjan Dey, Techno India College of Technology, Kolkata. We are thankful to the delegates and the authors for their participation and their interest in AICTC-2019 as a platform to share their ideas and innovation. We are also thankful to Dr. Amit Joshi, Director, Global Knowledge Research Foundation and Mr. Aninda Bose, Senior Editor, Hard Sciences, Springer, India, for providing continuous guidance and support. Also, we extend our heartfelt gratitude and thanks to the reviewers and technical program committee members for showing

their concern and effort during the review process. We are indeed thankful to everyone directly or indirectly associated with the conference organizing team, for leading it toward success. We hope you enjoy the conference proceedings and wish you all the best.

Bikaner, Rajasthan, India

Organizing Committee  
AICTC-2019

# Contents

<b>Blockchain Integrated Secured Scenarios in Advanced Wireless Networks .....</b>	1
Kailash Aseri	
<b>Analysis of Docker Performance in Cloud Environment .....</b>	9
Deepika Saxena and Navneet Sharma	
<b>A Review of Metaheuristic Techniques for Solving University Course Timetabling Problem .....</b>	19
Manpreet Kaur and Sanjay Saini	
<b>Using Social Media Analytics to Predict Social Media Engagement Outcome for Fortune CEOs .....</b>	27
Hitesha Yadav, Arpan K. Kar, and Smita Kashiramka	
<b>Smart Heart Attack Forewarning Model Using MapReduce Programming Paradigm .....</b>	37
Arushi Jain, Vishal Bhatnagar, and Annavarapu Chandra Sekhara Rao	
<b>Tuning of CNN Architecture by CSA for EMNIST Data .....</b>	45
Navdeep Bohra and Vishal Bhatnagar	
<b>Efficient Emergency Message DHC Broadcasting in Vehicular Ad Hoc Networks .....</b>	57
Jaipal, Dhanroop Mal Nagar, and Vinay Baghela	
<b>Software Effort Estimation Using Machine Learning Techniques .....</b>	65
Ripu Ranjan Sinha and Rajani Kumari Gora	
<b>Sentiment Analysis of English-Punjabi Code-Mixed Social Media Content to Predict Elections .....</b>	81
Mukhtiar Singh, Vishal Goyal, and Sahil Raj	

<b>Automatic Understanding of Code Mixed Social Media Text: A State of the Art . . . . .</b>	91
Neetika, Vishal Goyal, and Simpel Rani	
<b>Secure Server Virtualization Using Object Level Permission Model . . . . .</b>	101
Varsha Grover and Gagandeep	
<b>Implementing Slowloris DoS Using Docker . . . . .</b>	109
Ishaan Sharma, Manohit, and Abhinav Bhandari	
<b>Sentiment Analysis of Pulwama Attack Using Twitter Data . . . . .</b>	119
Ranu Lal Chouhan	
<b>A Survey on Architecture and Protocols for Wireless Sensor Networks . . . . .</b>	127
Anita Chandel, Vikram Singh Chouhan, and Dhawal Vyas	
<b>A Survey on Routing Protocols for Wireless Sensor Networks . . . . .</b>	143
Anita Chandel, Vikram Singh Chouhan, and Sunil Sharma	
<b>Drive into Future World Using Artificial Intelligence with Its Application in Sensor-Based Car Without Driver . . . . .</b>	165
Ridhima Sehgal	
<b>Linking and Digital Story Telling Approach in Teaching Towards Enhancing and Engagement of Smart Study . . . . .</b>	173
Sanjay Tejasvee and Manoj Kuri	
<b>Classifying Titanic Passenger Data and Prediction of Survival from Disaster . . . . .</b>	181
Shashank Shekhar, Deepak Arora, and Puneet Sharma	
<b>Soft Skills: An Integral Part of Technical Education . . . . .</b>	189
Nisha Srivastava and Manoj Kuri	
<b>Virtual Machine Migration Approach in Cloud Computing Using Genetic Algorithm . . . . .</b>	195
Gursharanjit Kaur and Rajan Sachdeva	
<b>A Survey on Electronic Health Records Using Cloud Computing Environment . . . . .</b>	205
Vivek Gehlot, S. P. Singh, and Akash Saxena	
<b>IoT Security Architecture with TEA for DoS Attacks Prevention . . . . .</b>	215
Vishal Sharma and Anand Sharma	
<b>Comparative Study of SVM and Naïve Bayes for Mangrove Detection Using Satellite Image . . . . .</b>	227
Anand Upadhyay, Santosh Singh, Nirbhay Singh, and Ajay Kumar Pal	

<b>Identification and Assessment of Black Sigatoka Disease in Banana Leaf . . . . .</b>	237
Anand Upadhyay, Neha Maria Oommen, and Siddhi Mahadik	
<b>Water Resource Detection Using High Resolution Satellite Image and GRNN . . . . .</b>	245
Anand Upadhyay, Manisha Pandey, and Ajay Kumar Pandey	
<b>Retinopathy Detection Using Probabilistic Neural Network . . . . .</b>	253
Anand Upadhyay, Parth Kantelia, and Rohan Parmar	
<b>Application of Unscented Kalman Filter for Parameter Estimation of Nonlinear Systems . . . . .</b>	261
Urmila Solanki, Ganesh P. Prajapat, and Manoj Chhimpa	
<b>Question Answering System Using LSTM and Keyword Generation . . . . .</b>	271
Minakshi Tomer and Manoj Kumar	
<b>Classification of LISS-III Image Using Fuzzy Logic . . . . .</b>	283
Anand Upadhyay, Sonam Mishra, and Aishwarya Khavadkar	
<b>Optimized Text Classification Using Deep Learning . . . . .</b>	293
Neeti Sangwan and Vishal Bhatnagar	
<b>Digital Learning: A Proficient Digital Learning Technology Beyond to Classroom and Traditional Learning . . . . .</b>	303
Sanjay Tejasvee, Devendra Gahlot, Rakesh Poonia, and Manoj Kuri	
<b>Data Security &amp; Future Issues for Cloud Computing . . . . .</b>	313
Devendra Gahlot, Sanjay Tejasvee, Kunal Bhushan Ranga, and Rishi Raj Vyas	
<b>Weather Event Prediction Using Combination of Data Mining Algorithms . . . . .</b>	319
Yogesh Kumar Jakhar, Nidhi Mishra, and Rakesh Poonia	
<b>Data Compression and Visualization Using PCA and T-SNE . . . . .</b>	327
Jyoti Pareek and Joel Jacob	
<b>Appscrumfall: APP Development Methodology Based on ScrumFall . . . . .</b>	339
Prerna Bisaa	
<b>Multiple Sequence Alignment Algorithm Using Adaptive Evolutionary Clustering . . . . .</b>	349
Jyoti Lakhani, Ajay Khunteta, Anupama Chowdhary, and Dharmesh Harwani	
<b>Theft Security System for Automatic Teller Machines Using IoT . . . . .</b>	365
Vinay Verma, Anjali Verma, Gaurav Sharma, and Anand Sharma	

<b>Tree-Based Multi-Keyword Rank Search Scheme Supporting Dynamic Update and Verifiability upon Encrypted Cloud Data . . . . .</b>	375
Pawan Kumar Tanwar, Ajay Khunteta, Vishal Goar, and Manoj Kuri	
<b>Techniques, Applications, and Issues in Mining Large-Scale Text Databases . . . . .</b>	385
Sandhya Avasthi, Ritu Chauhan, and Debi P. Acharjya	
<b>Vehicle Number Extraction Using Open Source Tools . . . . .</b>	397
Chetan Pandey, Amit Juyal, and Ankur Dumka	
<b>Classification of Energy Efficiency in Mobile Cloud Computing . . . . .</b>	409
Shubham Pal and Ankur Dumka	
<b>Perspectives of Blockchain in the Education Sector Pertaining to the Student's Records . . . . .</b>	419
Poonam Verma and Ankur Dumka	
<b>Ground-Level Water Prediction Using Time Series Statistical Model . . . . .</b>	427
Sandeep Kumar Mittal, Mamta Mittal, and Muhammad Sajjad Ali Khan	
<b>Prediction of Air Quality Index Using Hybrid Machine Learning Algorithm . . . . .</b>	439
Jasleen Kaur Sethi and Mamta Mittal	
<b>Employing Blockchain in Rice Supply Chain Management . . . . .</b>	451
M. Vinod Kumar, N. Ch. Sriman Narayana Iyengar, and Vishal Goar	
<b>Supervised Learning Method and Neural Network Algorithm for the Analysis of Diabetic Mellius and its Comparative Analysis . . . . .</b>	463
J. Jayashree, J. Vijayashree, N. Ch. Sriman Narayana Iyengar, and Vishal Goar	
<b>Nipah Virus Using Restricted Boltzmann Machine . . . . .</b>	477
Velpula Sandhya Rani, Havalath Balaji, Vishal Goar, and N. Ch. Sriman Narayana Iyengar	
<b>Big Data Analytics—Analysis and Comparison of Various Tools . . . . .</b>	491
Amit Gupta, Bhanu Prakash Dubey, Himani Sivaraman, and M. C. Lohani	
<b>Copy-Move Forgery Detection Methods: A Critique . . . . .</b>	501
Monika Kharanghar and Amit Doegar	
<b>Improving Website by Analysis of Web Server Logs Using Web Mining Tools . . . . .</b>	525
Neeraj Kandpal, Devesh Kumar Bandil, and M. S. Shekhawat	

Contents	xxi
----------	-----

<b>A New Approach for Paddy Leaf Blast Disease Prediction Using Logistic Regression .....</b>	533
Sree Charitha Kodat and Balaji Halavath	
<b>Assistive Technology for Students with Visual Impairments: A Resource for Teachers, Parents, and Students .....</b>	543
Amit Sadh	

# Editors and Contributors

## About the Editors

**Dr. Vishal Goar** holds M.C.A. degrees from Indira Gandhi National Open University, New Delhi, India, and a Ph.D. degree in Computer Science from SGVU University, Jaipur. He is currently an Assistant Professor at the Government Engineering College, Bikaner, where he is also Coordinator of the Research & Development Department. His research interests include various fields of computer programming, networking, databases, operating systems, cloud computing data mining. He has also organized several conferences and workshops, has delivered numerous lectures on technical innovations, and has published research papers at various conferences and in respected journals.

**Dr. Manoj Kuri** holds Bachelor's and Master of Engineering degrees in Electronics and Communication, a PG Diploma in Advanced Computing (CDAC Pune), and a Ph.D. from IIT Roorkee. He is currently an Assistant Professor and Head (ECE) of the Government Engineering College Bikaner, Rajasthan (India). He has 19 years of teaching and research experience and has published papers in various respected journals and conferences. He is a member of several scientific and professional societies, such as IEEE, ISTE and ISRS. His research interests include digital image processing, wireless sensor networks and microwave satellite SAR interferometry.

**Dr. Rajesh Kumar** holds a Bachelor of Engineering (honors) in Electrical Engineering from the National Institute of Technology, Kurukshetra, India; and a Master of Engineering (honors) in Power Engineering and a Ph.D. from Malaviya National Institute of Technology (MNIT), Jaipur, India, where he is currently a Professor at the Department of Electrical Engineering. He has also been a Research

Fellow at the National University of Singapore. His research focuses on computational intelligence, artificial intelligence, power & energy management, robotics, bioinformatics, smart grids and computer vision.

**Professor Tomonobu Senju** received his B.S. and M.S. degrees in Electrical Engineering from the University of the Ryukyus, Nishihara, Japan, in 1986 and 1988, respectively, and his Ph.D. degree in Electrical Engineering from Nagoya University, Nagoya, Japan, in 1994. He is currently a Full Professor at the Department of Electrical & Electronics Engineering, University of the Ryukyus. His research interests are in the areas of renewable energy, power system optimization, and operation and advanced control of electrical machines.

## Contributors

**Debi P. Acharjya** Vellore Institute of Technology, Vellore, India

**Deepak Arora** Department of Information Technology, Amity School of Engineering and Technology, Amity University, Lucknow Campus, Lucknow, India

**Kailash Aseri** Jodhpur National University, Rajasthan, India

**Sandhya Avasthi** Amity University, Noida, India

**Vinay Baghela** Software Engineering, Government Engineering College Bikaner, Bikaner, India

**Havalath Balaji** Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India;  
Government Engineering College, Bikaner, India

**Devesh Kumar Bandil** Suresh Gyan Vihar University, Jaipur, Rajasthan, India

**Abhinav Bhandari** Punjabi University, Patiala, India

**Vishal Bhatnagar** Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, India

**Prerna Bisaa** Tantia University, ShriGanganagar, India

**Navdeep Bohra** USICT, GGSIPU, New Delhi, India;  
Maharaja Surajmal Institute of Technology, New Delhi, India

**Anita Chandel** Information Technology Department, Engineering College, Bikaner, Bikaner, India

**Ritu Chauhan** Amity University, Noida, India

**Manoj Chhimpa** Department of Electrical Engineering, Government Engineering College, Bikaner, Rajasthan, India

**Ranu Lal Chouhan** Government Engineering College, Bikaner, Bikaner, India

**Vikram Singh Chouhan** Information Technology Department, Engineering College, Bikaner, Bikaner, India

**Anupama Chowdhary** Department of Computer Science, Keen College, Bikaner, India

**Amit Doegar** Computer Science & Engineering Department, NITTTR, Chandigarh, India

**Bhanu Prakash Dubey** Department of CSE, Graphic Era Hill University, Dehradun, India

**Ankur Dumka** Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India

**Gagandeep** Department of Computer Science, Punjabi University, Patiala, India

**Devendra Gahlot** MCA Department, Government Engineering College Bikaner, Bikaner, Rajasthan, India

**Vivek Gehlot** Research Scholar, Department of Computer Science and Engineering, Nims Institute of Engineering and Technology, Nims University Rajasthan, Jaipur, India

**Vishal Goar** Department of CA, Government Engineering College Bikaner, Bikaner, Rajasthan, India;

Sreenidhi Institute of Technology and Science, Hyderabad, Telangana, India

**Rajani Kumari Gora** Rajasthan Technical University, Kota, India;  
Computer Science, DCE, GoR, Jaipur, India

**Vishal Goyal** Department of Computer Science, Punjabi University, Patiala, Punjab, India

**Varsha Grover** Department of Computer Science, Punjabi University, Patiala, India

**Amit Gupta** Department of CSE, Graphic Era Hill University, Dehradun, India

**Balaji Halavath** Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, Telangana, India

**Dharmesh Harwani** Department of Microbiology, Maharaja Ganga Singh University, Bikaner, India

**Joel Jacob** Department of Computer Science, Gujarat University, Ahmedabad, India

**Arushi Jain** Indian Institute of Technology, Dhanbad, India

**Jaipal** Software Engineering, Government Engineering College Bikaner, Bikaner, India

**Yogesh Kumar Jakhar** Department of Computer Engineering, Poornima University, Jaipur, India

**J. Jayashree** School of Computer Science and Engineering, VIT, Vellore, Tamil Nadu, India

**Amit Juyal** Graphic Era Hill University, Dehradun, India

**Neeraj Kandpal** Suresh Gyan Vihar University, Jaipur, Rajasthan, India

**Parth Kantelia** Thakur College of Science and Commerce, Kandivali (E), Mumbai, India

**Arpan K. Kar** Department of Management Studies, Indian Institute of Technology, Delhi, India

**Smita Kashiramka** Department of Management Studies, Indian Institute of Technology, Delhi, India

**Gursharanjit Kaur** Guru Gobind Singh College of Modern Technology, Kharar, Punjab, India

**Manpreet Kaur** Department of Physics and Computer Science, Dayalbagh Educational Institute, Agra, India

**Muhammad Sajjad Ali Khan** Department of Mathematics, Institute of Numerical Sciences, Kohat University of Science and Technology, Kohat, Khyber Pakhtunkhwa, Pakistan

**Monika Kharanghar** Computer Science & Engineering Department, NITTTR, Chandigarh, India

**Aishwarya Khavadkar** Department of Information Technology, Thakur College of Science and Commerce, Kandivali (E), Mumbai, Maharashtra, India

**Ajay Khunteta** Department of Computer Engineering, Poornima University, Jaipur, India

**Sree Charitha Kodaty** Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, Telangana, India

**Manoj Kumar** Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India

**Manoj Kuri** Department of Electronics and Communication, Government Engineering College Bikaner, Bikaner, Rajasthan, India

**Jyoti Lakhani** Department of Computer Engineering, Poornima University, Jaipur, India;

Department of Computer Science, Maharaja Ganga Singh University, Bikaner, India

**M. C. Lohani** Department of CSE, Graphic Era Hill University, Dehradun, India

**Siddhi Mahadik** Department of Information Technology, Thakur College of Science and Commerce, Kandivali (E), Mumbai, India

**Manohit** Bachelor of Technology, Punjabi University, Patiala, India

**Nidhi Mishra** Poornima University, Jaipur, India

**Sonam Mishra** Department of Information Technology, Thakur College of Science and Commerce, Kandivali (E), Mumbai, Maharashtra, India

**Mamta Mittal** Department of Computer Science & Engineering, G. B. Pant Government Engineering College, New Delhi, India

**Sandeep Kumar Mittal** Department of Mathematics, G. B. Pant Government Engineering College, New Delhi, India

**Dhanroop Mal Nagar** Information Technology, Government Engineering College Bikaner, Bikaner, India

**Neetika** Department of Computer Science, College of Engineering & Management, Rampura Phul, Punjab, India

**Neha Maria Oommen** Department of Information Technology, Thakur College of Science and Commerce, Kandivali (E), Mumbai, India

**Ajay Kumar Pal** Thakur College of Science & Commerce, Kandivali (E), Mumbai, India

**Shubham Pal** Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India

**Ajay Kumar Pandey** Thakur College of Science and Commerce, Kandivali (E), Mumbai, India

**Chetan Pandey** Graphic Era Hill University, Dehradun, India

**Manisha Pandey** Thakur College of Science and Commerce, Kandivali (E), Mumbai, India

**Jyoti Pareek** Department of Computer Science, Gujarat University, Ahmedabad, India

**Rohan Parmar** Thakur College of Science and Commerce, Kandivali (E), Mumbai, India

**Rakesh Poonia** MCA Department, Government Engineering College Bikaner, Bikaner, Rajasthan, India

**Ganesh P. Prajapat** Department of Electrical Engineering, Government Engineering College, Bikaner, Rajasthan, India

**Sahil Raj** School of Management and Studies, Punjabi University, Patiala, India

**Kunal Bhushan Ranga** MCA Department, Government Engineering College Bikaner, Bikaner, Rajasthan, India

**Simpel Rani** Department of Computer Science and Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

**Velpula Sandhya Rani** Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India;  
Government Engineering College, Bikaner, India

**Annavarapu Chandra Sekhara Rao** Indian Institute of Technology, Dhanbad, India

**Rajan Sachdeva** Guru Gobind Singh College of Modern Technology, Kharar, Punjab, India

**Amit Sadh** Department of Secondary Education, Bikaner, Rajasthan, India

**Sanjay Saini** Department of Physics and Computer Science, Dayalbagh Educational Institute, Agra, India

**Neeti Sangwan** GGS Indraprastha University, New Delhi, India;  
MSIT, New Delhi, India

**Akash Saxena** Professor, Department of Computer Science, Compucom Institute of Information Technology and Management, Jaipur, India

**Deepika Saxena** Computer Science, The IIS University, Jaipur, India

**Ridhima Sehgal** Computer Science, BBK DAV College, Amritsar, India

**Jasleen Kaur Sethi** University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, New Delhi, India

**Anand Sharma** CSE, SET, Mody University of Science and Technology, Lakshmangarh, Sikar, India

**Gaurav Sharma** Jaipur National University, Jaipur, India

**Ishaan Sharma** Bachelor of Technology, Punjabi University, Patiala, India

**Navneet Sharma** Computer Science, The IIS University, Jaipur, India

**Puneet Sharma** Department of Information Technology, Amity School of Engineering and Technology, Amity University, Lucknow Campus, Lucknow, India

**Sunil Sharma** Computer Science Department, Government Polytechnic College, Jhalawar, Jhalawar, India

**Vishal Sharma** CSE, SET, Mody University of Science and Technology, Lakshmangarh, India

**Shashank Shekhar** Department of Information Technology, Amity School of Engineering and Technology, Amity University, Lucknow Campus, Lucknow, India

**M. S. Shekhawat** Department of Physics, Government Engineering College, Bikaner, Rajasthan, India

**Mukhtiar Singh** Department of Computer Science, Punjabi University, Patiala, India

**Nirbhay Singh** Thakur College of Science & Commerce, Kandivali (E), Mumbai, India

**S. P. Singh** Professor and Head, Department of Computer Science and Engineering, Nims Institute of Engineering and Technology, Nims University Rajasthan, Jaipur, India

**Santosh Singh** Thakur College of Science & Commerce, Kandivali (E), Mumbai, India

**Ripu Ranjan Sinha** Rajasthan Technical University, Kota, India;  
SS Jain Subodh College, Jaipur, India

**Himani Sivaraman** Department of CSE, Graphic Era Hill University, Dehradun, India

**Urmila Solanki** Department of Electrical Engineering, Government Engineering College, Bikaner, Rajasthan, India

**N. Ch. Sriman Narayana Iyengar** Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India;  
Government Engineering College, Bikaner, India

**Nisha Srivastava** Engineering College Bikaner, Bikaner, India

**Pawan Kumar Tanwar** Poornima University, Jaipur, India;  
Engineering College Bikaner, Bikaner, India

**Sanjay Tejasvee** MCA Department, Government Engineering College Bikaner, Bikaner, Rajasthan, India

**Minakshi Tomer** University School of Information Communication and Technology, GGSIPU, Delhi, India;  
Maharaja Surajmal Institute of Technology, Delhi, India

**Anand Upadhyay** Department of Information Technology, Thakur College of Science and Commerce, Kandivali (E), Mumbai, Maharashtra, India

**Anjali Verma** Banasthali Vidyapeeth, Tonk, India

**Poonam Verma** Graphic Era Deemed to be University, Dehradun, India

**Vinay Verma** Mody University of Science and Technology, Sikar, India

**J. Vijayashree** School of Computer Science and Engineering, VIT, Vellore, Tamil Nadu, India

**M. Vinod Kumar** Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India

**Dhawal Vyas** Information Technology Department, Engineering College, Bharatput, Bharatput, India

**Rishi Raj Vyas** CSE Department, Government Engineering College Bikaner, Bikaner, Rajasthan, India

**Hitesha Yadav** Department of Management Studies, Indian Institute of Technology, Delhi, India

# Blockchain Integrated Secured Scenarios in Advanced Wireless Networks



Kailash Aseri

**Abstract** The wireless networks are nowadays quite susceptible towards the assorted assaults, and thereby the needs arise to secure the overall scenarios. The Blockchain technology is one of the prominent and high-performance approaches that can be used for the integration of security with wireless networks to enforce a greater degree of security and overall performance. Blockchain refers to the high-performance and security-aware technology in which a digital ledger is maintained. The digital ledger is quite transparent, and there is no scope of any manipulations in the records by the intermediates or any administrator. The records of all the transactions are logged in the Blockchain ledger, and the operations are committed finally with different protocols and algorithms which cannot be hacked by the third-party intrusions.

**Keywords** Blockchain and wireless networks · Security-aware wireless networks · Wireless networks with blockchain

## 1 Introduction

In advanced wireless scenarios, the overall performance is very important and required so that the overall network environment can be made secured [1, 2]. The work is presenting the usage of Blockchain-based implementations with the wireless networks using Python-based libraries which can be integrated with Raspberry Pi or Arduino or any other open-source board for the integration and enforcement of scenarios [3–5].

Blockchain is the state-of-the-art technology that is always associated with security and higher degree of privacy in assorted applications [6–8]. Nowadays, Blockchain technology is not limited to the cryptocurrencies, rather it is under implementation for various social and corporate segments. These segments include

---

K. Aseri (✉)  
Jodhpur National University, Rajasthan, India  
e-mail: [kailash.aseri@gmail.com](mailto:kailash.aseri@gmail.com)

e-governance, social networking, e-commerce, transportation, logistics, professional communications and many others [9, 10].

Following are some of the examples of Blockchain implementations:

Entertainment: KickCity B2Expand Spotify Guts Veredictum

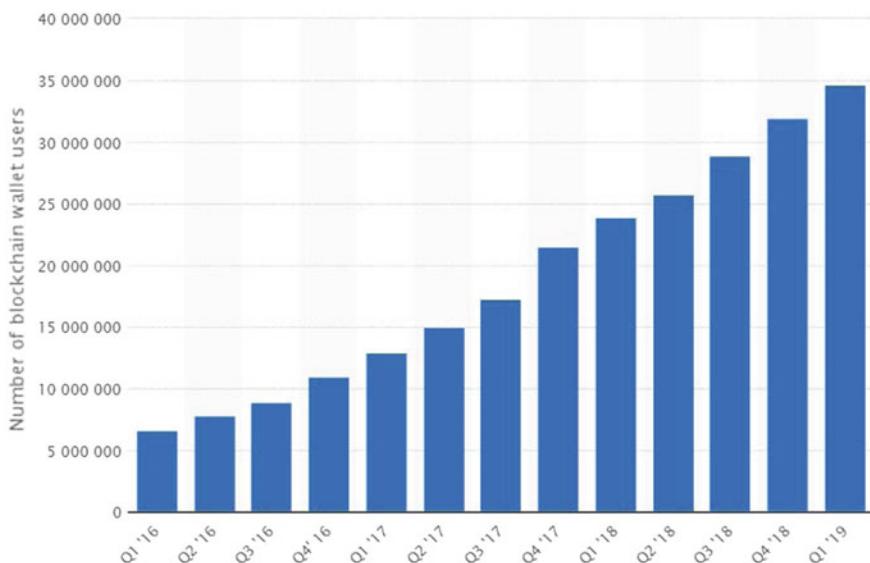
Social Networks: Matchpool Minds MeWe Steepshot DTube Mastodon Sola

Cryptocurrency: Bitcoin Litecoin Namecoin Dogecoin Primecoin Nxt Ripple Ethereum

Retail: Warranteer Blockpoint Loyyal Fluz Fluz Shopin Spl.yt Opskins ECoin-merce.io Every.Shop Portion Buying.com (Fig. 1).

In Blockchain network, there exist blocks of different data elements and records. Each block participates in the Blockchain network, and it is immutable. The term immutable here refers that it is secured and non-breakable. Hence, it forms the Blockchain of secured chain of blocks without any probabilities of intentional or accidental tampering or leakage in the data [11, 12].

The first block in the chain of network is known as the genesis block from where the Blockchain initiates the transactions. With the increase by inserting different blocks with encryption of every block with the previous block, it becomes secured and therefore difficult to crack the previous states because of so many encryptions [13–15].



**Fig. 1** Number of Blockchain wallet users worldwide (2016–2019) (Source Statista, the Statistics Portal)

## 2 Wireless Network with Python-Based Blockchain Programming

Python is the key programming language that is used in almost every area of high-performance computations. Python provides the tools and libraries which can be used for the Blockchain development including decentralized applications. As in Blockchain technology, there are secured protocols and algorithms, and the Python programming is having enormous toolkits available on its official repository <https://pypi.org/>. The specific libraries for decentralized applications and Blockchain implementations are available at the following URLs

- <https://pypi.org/search/?q=blockchain>
- <https://pypi.org/search/?q=dapp>.

The additional libraries and toolkits can be installed with Raspberry Pi or Arduino with existing Python using pip installer.

The Blockchain technology is highly dependent and making use on the integrations with dynamic cryptography and encryption. For this, the hashlib library can be installed using the above-mentioned instruction.

## 3 Building Blockchain Applications

Following is the scenario of a secured Blockchain that is generating the hash values so that overall transactions and records will be highly secured. In the following code, the dynamic hash value is generated that is the base of any Blockchain with different transactions in a chain and that makes the overall Blockchain.

```

class Block:
    def __init__(self, idx, ts, mydata, backhash):
        self.idx = idx
        self.ts = ts
        self.mydata = mydata
        self.backhash = backhash
        self.hash = self.hashop()

    def hashop(self):
        shashash = hasher.sha256()
        shashash.update(str(self.idx) + str(self.ts) + str(self.mydata) + str(self.backhash))
        return shashash.hexdigest()

    def genesis():
        return Block(0, date.datetime.now(), "Genesis Block", "0")

    def next_block(last_block):
        this_idx = last_block.idx + 1
        this_ts = date.datetime.now()
        this_mydata = "Block" + str(this_idx)
        this_hash = last_block.hash
        return Block(this_idx, this_ts, this_mydata, this_hash)

    blockchain = [genesis()]
    back_block = blockchain[0]
    maxblocks = 20

    for i in range(0, maxblocks):
        block_to_add = next_block(back_block)
        blockchain.append(block_to_add)
        back_block = block_to_add
        print "Block #{} inserted in Blockchain".format(block_to_add.idx)
        print "Hash Value: {}".format(block_to_add.hash)

```

With the execution of code, the following outcome is obtained that is having different hash values and provides higher degree of security using cryptography functions. Using these hash values, the attempt of hacking or sniffing the transaction will be almost impossible (Fig. 2).

```

E:\Python27\blockchain>python blockchainhash.py
Block #1 inserted in Blockchain
Hash Value: e7de0e16bd31d89de438f3034b744dc10f1eea4adff948960e0828914d0f7b66

Block #2 inserted in Blockchain
Hash Value: 919225537d90c280e04dc9d344389789d359e97b12a28069729db2c256b4422e

Block #3 inserted in Blockchain
Hash Value: 377247057a84c0ba35c1ea0196fefcc8660ee9ac5dcbbd4b424411ffc3d224f2

Block #4 inserted in Blockchain
Hash Value: 7755fe827549ac829c8d509783d3fafd2e4fe019afca8ea8ce8bb84014c1f9c1

Block #5 inserted in Blockchain
Hash Value: 707d280e76a49a6576b6016cfac4a667cc42bd90f6448858df2dabdc831086a7

Block #6 inserted in Blockchain
Hash Value: 1738e933f8a662141c258ead3fdb6a55cb18281169901653667f135c855e0c10

```

**Fig. 2** Generation of hash values using blockchain implementation

## 4 Deployment of Wireless Network-Based Distributed Blockchains

As in the earlier example, the implementation of hash function with the blocks is done on stand-alone system. In case of actual Blockchain, it is required to be distributed so that different users can initiate their transactions and blocks.

For distributed and web-based implementations, there are different frameworks in Python.

In Blockchain programming, the Proof-of-Work (PoW) is one of the very important algorithms. It is used to confirm and validate the transactions so that the new blocks are added in the Blockchain. It is referred as the key consensus algorithm for the verification and authenticity of the transactions. In Blockchain network, different miners participate for the validation, completing the transactions. For the successful validations, the miners are rewarded with the digital cryptocurrencies as their remuneration [16–18].

This process also avoids the double spending problem so that the digital currency or transaction is implemented in a secured way. For example, if A transmits a file or wireless message to B, in this case, that specific file or currency values in the records of A must be deleted and then should be reflected in the records of B. Traditionally, it is done by the wireless controller as intermediate. In case of Blockchain network, it is implemented without any intermediate, and it is validated automatically using specialized algorithms [19, 20]. If there are instances of not deleting the transaction from the sender, it will de-evaluate the wireless message despite the type of wireless message.

```
miner_address = "*****"
myblockchain = []
myblockchain.append(create_genesis_block())
this_nodes_transactions = []
peer_nodes = []
mining = True
@node.route('/myblockchain', methods=['POST'])
def transaction():
    new_myblockchain = request.get_json()
```

```

print "Amount: {}\n".format(new_myblockchain['amount'])
return "Transaction Successful\n"
@node.route('/blocks', methods=['GET'])
def get_blocks():
    chain_to_send = myblockchain
    for i in range(len(chain_to_send)):
        block = chain_to_send[i]
        block_idx = str(block.idx)
        block_timestamp = str(block.timestamp)
        block_data = str(block.data)
        block_hash = block.hash
        chain_to_send[i] = {
            "idx": block_idx,
            "timestamp": block_timestamp,
            "data": block_data,
            "hash": block_hash
        }

```

Figure 3 depicts the web-based implementation of Blockchain using web server so that the distributed deployment shall be there.

Wireless API "`http://localhost:5000/myblockchain`" -d `{"from": "ss", "to": "fsd", "amount": 3}`" -H "Content-Type:application/json"

Using cURL, the transaction can be implemented, and its impact on the Blockchain will be visualized (Fig. 4).

The cURL library for Windows operating system can be installed from <https://curl.haxx.se/windows/> (Fig. 5).

As depicted in Fig. 6, the execution of code and overall implementation with all the records and transactions can be analysed so that the transparency of operations will be there without any attempt of hacking. Using Proof-of-Work (PoW), the integrity of transactions is logged and committed.

```

E:\>cd Python27
E:\Python27>cd blockchain
E:\Python27\blockchain>python blockchainserver.py
* Serving Flask app "blockchainserver" (lazy loading)
* Environment: production
WARNING: Do not use the development server in a production environment.
Use a production WSGI server instead.
* Debug mode: off
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)

```

**Fig. 3** Executing wireless board integrated Blockchain code

```
E:\curl-7.65.0-win64-mingw\bin>curl "http://localhost:5000/myblockchain" -d "{\"from\":\"Sender\",\"to\":\"Receiver\", \"amount\":5}" -H "Content-Type:application/json"
Transaction Successful

E:\curl-7.65.0-win64-mingw\bin>curl "http://localhost:5000/myblockchain" -d "{\"from\":\"Person-1\",\"to\":\"Person-2\", \"amount\":5}" -H "Content-Type:application/json"
Transaction Successful

E:\curl-7.65.0-win64-mingw\bin>
```

**Fig. 4** Performing the transaction on wireless chips

```
E:\curl-7.65.0-win64-mingw\bin>curl localhost:5000/mine
{"timestamp": "2019-05-25 20:05:42.338000", "data": {"transactions": [{"to": "Receiver", "amount": 5, "from": "Sender"}, {"to": "Person-2", "amount": 5, "from": "Person-1"}, {"to": "*****", "amount": 1, "from": "network"}, {"proof-of-work": 18}, "hash": "8ae9acf42e4c4b89384818b0bd1a1ce48640734fcc5f2813bfb4368eed47d4d", "idx": 1}]

E:\curl-7.65.0-win64-mingw\bin>
```

**Fig. 5** Mining the records and transactions on wireless controller

```
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
127.0.0.1 - - [25/May/2019 20:03:36] "POST /txion HTTP/1.1" 404 -
127.0.0.1 - - [25/May/2019 20:04:01] "POST /txion HTTP/1.1" 404 -
New transaction
Sender: Sender
Receiver: Receiver
Amount: 5

127.0.0.1 - - [25/May/2019 20:04:23] "POST /myblockchain HTTP/1.1" 200 -
New transaction
Sender: Person-1
Receiver: Person-2
Amount: 5
```

**Fig. 6** Recording of all transactions on network controller

## 5 Conclusion

In current scenarios, the governments as well as corporate organizations are striving towards the implementation of Blockchain technology for secured application. For these integrations, there is a need to associate the secured algorithms of Proof-of-Work (PoW) so that the privacy and integrity of implementations will be there. The research scholars and forensic scientists can make use of Blockchain technologies so that the exact and accurate prediction of specific identities shall be there which can be used for the criminal forensic as well as the law enforcement scenario.

## References

1. Dorri A, Steger M, Kanhere SS, Jurdak R (2017) Blockchain: a distributed solution to automotive security and privacy. *IEEE Commun Mag* 55(12):119–125
2. Khan MA, Salah K (2018) IoT security: review, blockchain solutions, and open challenges. *Future Gener Comput Syst* 82:395–411
3. Dorri A, Kanhere SS, Jurdak R, Gauravaram P (2017) Blockchain for IoT security and privacy: the case study of a smart home. In: 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, pp 618–623
4. Alphand O, Amoretti M, Claeys T, Dall'Asta S, Duda A, Ferrari G et al (2018) IoTChain: a blockchain security architecture for the Internet of Things. In: 2018 IEEE wireless communications and networking conference (WCNC). IEEE, pp 1–6
5. Liu H, Zhang Y, Yang T (2018) Blockchain-enabled security in electric vehicles cloud and edge computing. *IEEE Netw* 32(3):78–83
6. Guan Z, Si G, Zhang X, Wu L, Guizani N, Du X, Ma Y (2018) Privacy-preserving and efficient aggregation based on blockchain for power grid communications in smart communities. *IEEE Commun Mag* 56(7):82–88
7. Banerjee M, Lee J, Choo KKR (2018) A blockchain future for internet of things security: a position paper. *Digital Commun Netw* 4(3):149–160
8. Biswas K, Muthukumaraasamy V (2016) Securing smart cities using blockchain technology. In: 2016 IEEE 18th international conference on high performance computing and communications; IEEE 14th international conference on smart city; IEEE 2nd international conference on data science and systems (HPCC/SmartCity/DSS). IEEE, pp 1392–1393
9. Minoli D, Occhiogrosso B (2018) Blockchain mechanisms for IoT security. *Internet Things* 1:1–13
10. Park J, Park J (2017) Blockchain security in cloud computing: use cases, challenges, and solutions. *Symmetry* 9(8):164
11. Sharma V, You I, Kul G (2017) Socializing drones for inter-service operability in ultra-dense wireless networks using blockchain. In: Proceedings of the 2017 international workshop on managing insider security threats. ACM, pp 81–84
12. Dorri A, Kanhere SS, Jurdak R (2017) Towards an optimized blockchain for IoT. In: Proceedings of the second international conference on Internet-of-Things design and implementation. ACM, pp 173–178
13. Puthal D, Malik N, Mohanty SP, Kougnanos E, Yang C (2018) The blockchain as a decentralized security framework [future directions]. *IEEE Consum Electron Mag* 7(2):18–21
14. Kotobi K, Bilen SG (2018) Secure blockchains for dynamic spectrum access: a decentralized database in moving cognitive radio networks enhances security and user access. *IEEE Veh Technol Mag* 13(1):32–39
15. Joshi AP, Han M, Wang Y (2018) A survey on security and privacy issues of blockchain technology. *Math Found Comput* 1(2):121–147
16. Lasla N, Younis M, Znaidi W, Arbia DB (2018) Efficient distributed admission and revocation using blockchain for cooperative ITS. In: 2018 9th IFIP international conference on new technologies, mobility and security (NTMS). IEEE, pp 1–5
17. Wu L, Du X, Wang W, Lin B (2018) An out-of-band authentication scheme for internet of things using blockchain technology. In: 2018 international conference on computing, networking and communications (ICNC). IEEE, pp 769–773
18. Kumar NM, Mallik PK (2018) Blockchain technology for security issues and challenges in IoT. *Procedia Comput Sci* 132:1815–1823
19. Lee JH, Kim H (2017) Security and privacy challenges in the internet of things [security and privacy matters]. *IEEE Consum Electron Mag* 6(3):134–136
20. Esposito C, De Santis A, Tortora G, Chang H, Choo KKR (2018) Blockchain: a panacea for healthcare cloud-based data security and privacy? *IEEE Cloud Comput* 5(1):31–37

# Analysis of Docker Performance in Cloud Environment



Deepika Saxena and Navneet Sharma

**Abstract** In the present scenario, the technology is going very different, and this difference lies on different–different platforms. The platform is assured of reliability, consistency, and quickness. These bundles of quality are called container (Docker and LXC). Container helps to produce operational efficiency, version control, developer productivity, and environment consistency. The technical industry is adopting the container technology in both internal and commercial uses. In this paper, we analyze the performance of Docker by using different applications or tools in cloud environment.

**Keywords** Docker · Containers · Cloud computing · Hypervisor · Virtualization · Privacy · Virtual machine

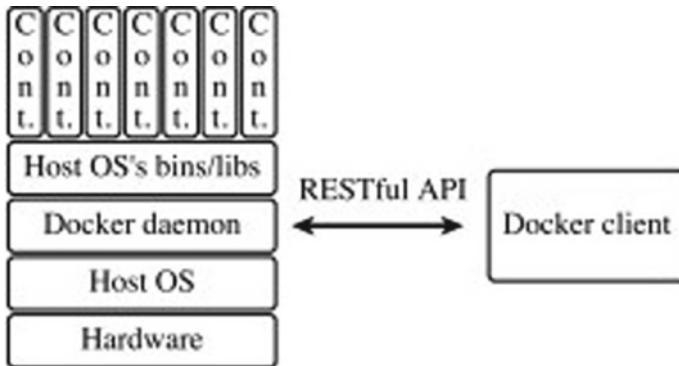
## 1 Introduction

Docker is a containerization platform that wraps our application and leaning together in the form of a Docker container to ensure that our exercise works polished in cloud environment. Docker has an ability to build, ship, and run the application in the cloud environment, because it is an open-source technology. It has many popular applications, such as Spotify, Yelp, and Ebay. Although container technologies have been famous in the present scenario, Docker has many controlling technologies. First, it provides coalition to simply and safely create and control containers. Second, developers can use the applications with Docker without any modification. Furthermore, Docker can create more virtual environments than other technologies on the same hardware [1]. Last but not least, Docker is well handled with third-party tools, like Calinciuc [2], preeth [3], and Menage [4]. They are integrated with

---

D. Saxena (✉) · N. Sharma  
Computer Science, The IIS University, Jaipur, India  
e-mail: [sunshine.deepika@gmail.com](mailto:sunshine.deepika@gmail.com)

N. Sharma  
e-mail: [navneet.sharma@iisuniv.ac.in](mailto:navneet.sharma@iisuniv.ac.in)



**Fig. 1** Architecture of Docker engine

Docker, so that they make Docker containers deploy easily in cloud environment. Besides, many orchestration tools, like Kathawala [5], Zhao [6], and Kubernetes [7], also support Docker containers. Docker has many components like: Docker engine, Docker images, and Docker hub.

## 1.1 *Docker Engine*

Docker engine is a portable tool [8] which is based on container virtualization. That is why, the Docker architecture is similar to the container (Fig. 1).

The Docker containers run on top of the Docker daemon which is managing all Docker containers. Here, Docker client provides the interface with container and Docker. They will accept the commands from the users and send it to Docker daemon. By using this command from the user, send it to Docker Daemon with the help of RESTful APIs. Using these communication techniques, Docker runs on the same and different host containers.

## 1.2 *Docker Hub*

Docker hub is repository images for both public and private areas. Users can share their images in the public or private sector. Docker has a duty to identify or recognize the images where their owner submitted into the hub.

### 1.3 Docker Images

Two methods will consider building an image. First is read-only template; every image has their base image. The base images are operating system images like Ubuntu and Fedora. Those images of OS will help to create a container with the ability of complete running OS. The second technique is to build a Docker file. That file has a list of instructions to create a Docker image.

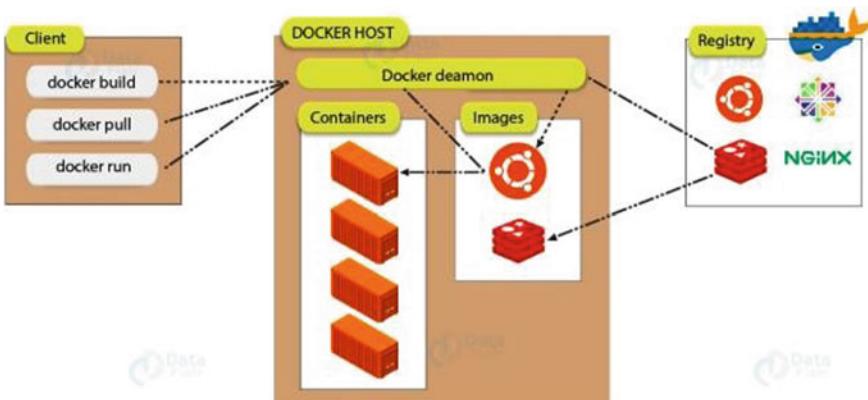
## 2 Internal Components of Docker

Three internal components are used in Docker: Docker client–server, Docker registers, and Docker container.

### 2.1 Docker Client–Server

Docker can be described by client–server-based application as shown in Fig. 2.

Basically, Docker host runs the Docker daemon, daemon takes Docker API request like: ‘Docker run,’ ‘Docker builds,’ and manages Docker objects. Client is the way where many Docker users interact with the Docker. The Docker client sends the commands like Docker build, Docker run to the Docker daemon, which carries it. Docker client has the ability to communicate more than one daemon.



**Fig. 2** Docker architecture

## 2.2 Docker Registers

Register is server-side application, which stores and distributes Docker images. Docker registry is a storage and distribution technique. A Docker registry is managed into Docker repositories. The registry gives the push and pulls the images for Docker and pushes the new images to the register.

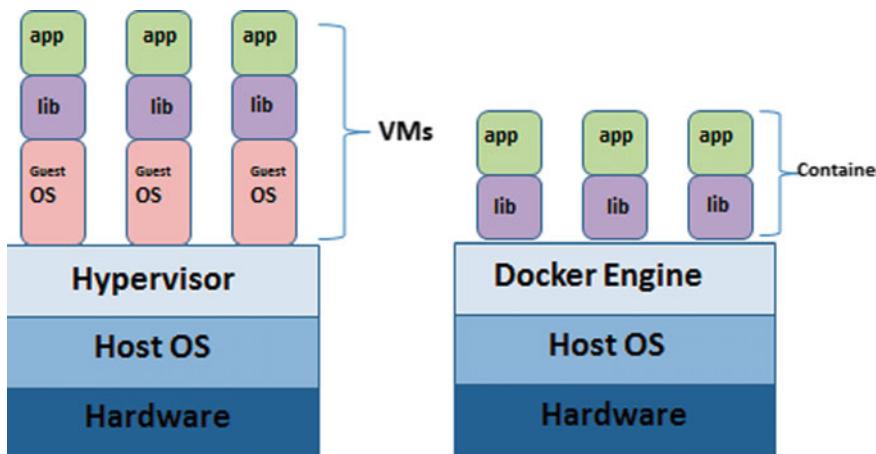
## 2.3 Docker Container

Docker container is a software that is based on the open-source concept. It packages the application in container and gives permission to migrate to any operating system.

## 3 Docker Versus Virtual Machine

Docker is based on containers, so Docker is container-based technology. And container has a quality to take space on OS. It aims for running applications. In Docker, container shares the host kernel. Virtual machine is not based on container. They manage or create user space of OS. Every VM has OS and application. In the case of virtual machine, each workload needs a complete OS, but with a container environment, multiple workload can run on one OS. Docker application runs on the system, and they does not require hypervisor (Fig. 3).

Your contribution should be prepared in Microsoft Word. Docker is a concept of developing, delivering, and running a platform of data or instructions. Docker is



**Fig. 3** Comparison diagram of Docker container and VM

a container administrative services provider. It is able to disjoint user applications from their infrastructure, and hence, it delivers software very fast. By this facility, it is possible to manage our infrastructure. Docker is a famous developer company that provides types of facilities with supports of cloud, Linux, and Windows vendors including Microsoft. Docker makes the process of application deployment very easy and efficient and resolves a lot of issues related to deploying application. Docker gives us a typical way to pack our application with all its dependencies in a container. Containers allow a developer to suite up an application with all of the parts it needs, such as libraries and other needs and ship it all out as one set. There are different environments on Windows, Linux, or Mac OS. On the development computer, the developers run a Docker host where Docker images are deployed. Docker provides a discrete image of a file system with everything the application requires during runtime [9]. Containers contain these libraries and tools that particular applications require. Developer who works on Linux or on the Mac uses a Docker host that is Linux-based, and they can create images separately only of Linux containers. So, by this, all Windows images can run only on Windows host and Linux images can run on Linux host.

### ***3.1 Related Operation***

Felter et al. give execution analysis on BareMetal, VM, and LXC. The result shows factors responsible for analyzed performance. The result is based on different-different platform. Their result shows that LXC causes some issues with VM, whereas BareMetal defines to faster concept than LXC. Define performance analyses for VM and Docker containers. The analysis part is based on CPU, memory disk, and system performance. Their evaluation shows that Docker container is better than the VM. Here, we performed a comparative analysis which is based on BareMetal, VM, and LXC. The analysis defines the variation between LXC and VM. Our research defines that LXC is good to use than VM. Ali Babar et al. measured the analysis of host OS, VM, and LXC. Their performance defines that LXC is good over VM. Yamato measured the start time in BareMetal, Docker, and VM in cloud environment. The result describes that the Docker container start time is very less against the VM [10]. There are various projects in the open stack cloud scenario that are trying to improve the performance and get better results.

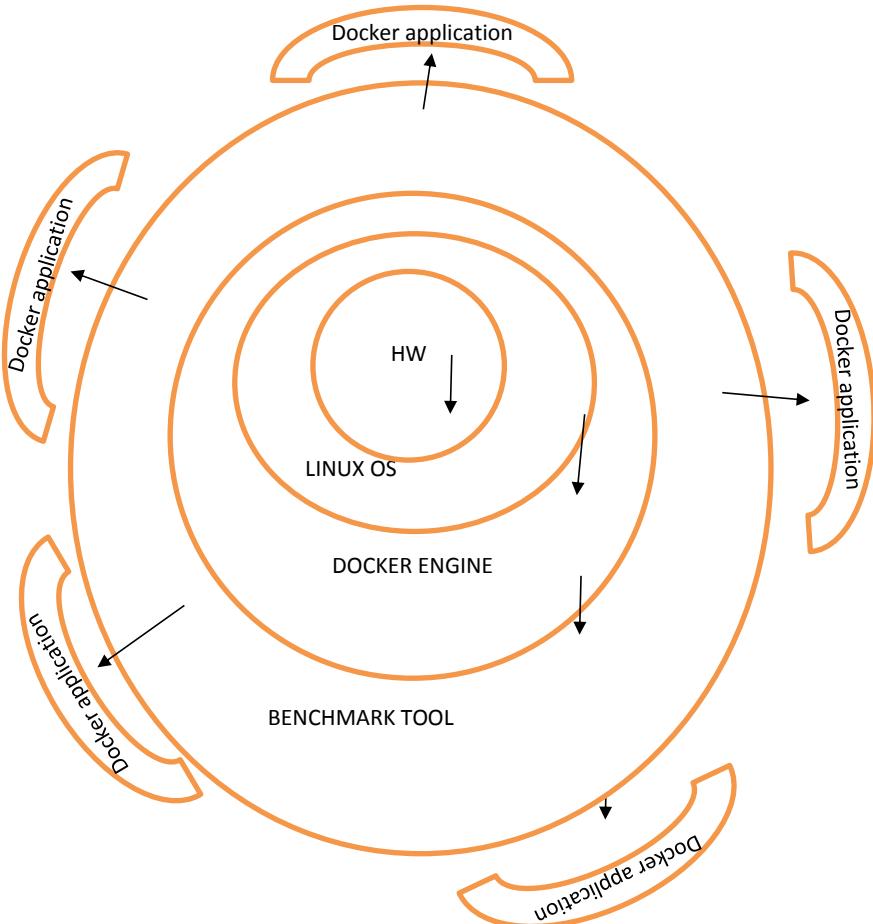
According to Fisher [11] lmbench and BYTE mark Native Mode Benchmark ver. 2(nbench), open-source benchmarks are to measure the conducting of each environment.

## 4 Planned Work

Docker container is used for calculating the start time that is needed for running an application.

So, the study is conducted to identify how the start time can be increased by using the benchmark tool. It will be very helpful for the developer to decide the platform on which the application will run.

Figure 4. Our analysis model is based on benchmarking Docker container with layered architecture. This model is based on both the virtual machine and environment of the BareMetal. The benchmark tool will gain the Docker engine in start time. Experiment calculation setup is based on Linux. The benchmark tool is to line up Apache Web server. The Apache Web services are efficient in calculating the start



**Fig. 4** Architecture of proposed system

time for each arrangement. To measure the performance, Web service is grouping in the count of 10, 20, 30, and 40.

## 5 Benchmark

Benchmark tool acts as a testing tool that measures the output performance of hardware. There are many types of benchmark like: component benchmark, synthetic benchmark, database benchmarks, I/O benchmarks, real program, and parallel benchmark [5].

## 6 Evaluation

Benchmark tool helps in studying the start time of Docker container application which acts as a base for BareMetal and VM. Start time signifies the time that is required for introducing app on Docker container. Images 4 is used for hosting the Docker container Lingayat [12]. The image acts as a very useful tool in maintaining the benchmark environment. Usually, start time is computed for BareMetal and average time is computed for VM. In order to test the 10, 20, 30, and 40, the outcomes are repeated ten times for every situation. Tables 1, 2, 3, and 4 shows the results of Web server displayed as container images for 10, 20, 30, and 40 application, respectively.

**Table 1** Result for 10 Docker images

	BareMetal	Virtual machine
Total time for ten results	165.001903770	275.10932001
Average time	16.5001903770	27.510932001

**Table 2** Result for 20 Docker images

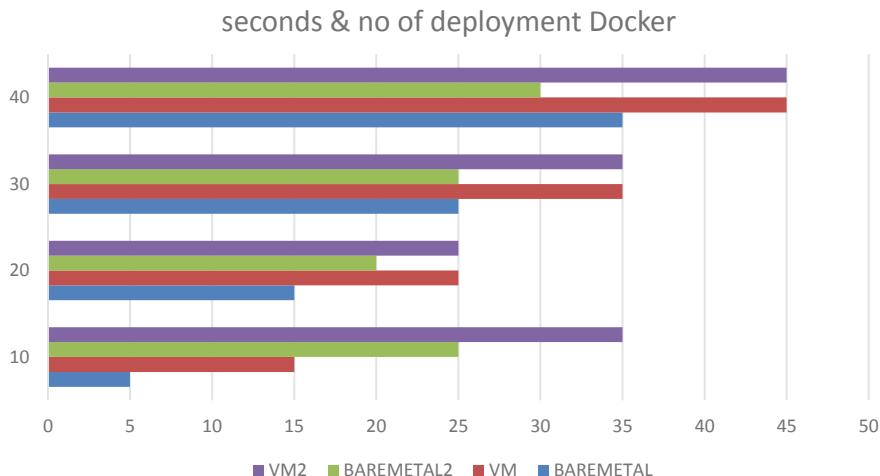
	BareMetal	Virtual machine
Total time for ten results	255.49190377	438.58490377
Average time	25.549190377	43.858490377

**Table 3** Result for 30 Docker images

	BareMetal	Virtual machine
Total time for ten results	348.313299256	551.68274745
Average time	34.8313299356	55.168274745

**Table 4** Result for 40 Docker images

	BareMetal	Virtual machine
Total time for ten results	437.875238256	673.281662807
Average time	43.7875238256	67.3281662907

**Fig. 5** Shows comparison between BareMetal and VM

By using the benchmark technique, the bar chart (Fig. 5) shows the comparison between BareMetal and VM. By the presence of hardware, the VM is very tardy in their start time. Docker container images are quicker in their start time as compared to VM. As a result, the count does not affect the performance of the Docker. However, the count affects the performance in VM.

## 7 Conclusion

This paper highlights the working of Docker container and its significance in the BareMetal environment. Virtual machine results were very leisurelier than BareMetal. This slower performance of the virtual machine is due to its architecture. In the comparison of both VM and BareMetal, the performance shows that BareMetal is faster than VM, its very time-saving concept, memory, and size-saving concept. To get better results, it is essential to arrange Docker container on BareMetal machines.

## References

1. Yamato Y (2015) OpenStack hypervisor, container and baremetal servers performance comparison. *IEICE Commun Express* 4(7):228–232
2. Fisher M (2012) Performance benchmarking physical and virtual linux environments. MA thesis, University of Cape Town
3. Caliniciuc A, Spoiala CC, Turcu CO, Filote C (2016) OpenStack and docker: building a high-performance IaaS platform for interactive social media applications. In: 2016 international conference on development and application systems (DAS), May 2016. IEEE
4. Harter T, Salmon B, Liu R, Arpacı-Dusseau AC, Arpacı-Dusseau RH (2016) Slacker: fast distribution with lazy docker containers. In: 14th USENIX conference on file and storage technologies (FAST 16). USENIX Association, Santa Clara, CA, pp 181–195 [Online]. Available: <https://www.usenix.org/conference/fast16/technicalsessions/presentation/harter>
5. Kozhirkayev Z, Sinnott RO (2017) A performance comparison of container-based technologies for the cloud. *Future Gener Comput Syst* 68:175–182
6. Peinl R, Holzschuh F, Pfitzer F (2016) Docker cluster management for the cloud—survey results and own solution. *J Grid Comput* 14(2):265–282
7. Elmuti D, Kathawala Y (1997) An overview of benchmarking process: a tool for continuous improvement and competitive advantage. *Benchmarking Qual Manag Technol* 4(4):229–243
8. Ratan V (2017) Docker: a favourite in the devops world [Online]. Available: <http://www.opensourceforu.com/2017/02/dockerfavourite-devops-world/>
9. Preeth EN, Mulerickal FJP, Paul B, Sastri Y (2015) Evaluation of docker containers based on hardware utilization. In: 2015 international conference on control communication & computing India (ICCC), Nov 2015. IEEE
10. Shetty J, Upadhyaya S, Rajarajeshwari H, Shobha G, Chandra J (2017) An empirical performance evaluation of docker container, openstack virtual machine and bare metal server. *Indonesian J Electr Eng Comput Sci* 7(1):205–213
11. Ali Babar M, Ramsey B (2017) Evaluating docker for secure and scalable private cloud with container technologies. Technical report. CREST, University of Adelaide, Adelaide
12. Liu D, Zhao L (2014) The research and implementation of cloud computing platform based on docker. In: 2014 11th international computer conference on wavelet active media technology and information processing (ICCWAMTIP), Dec 2014. IEEE
13. Pahl C (2015) Containerization and the PaaS cloud. *IEEE Cloud Comput* 2(3):24–31
14. Lingayat A, Singh A, Naik V, Badre RR, Gupta AK (2018) Horizon, a web-based user interface for managing services in openstack: an introspection. In: 9th international conference on computing, communication and networking technologies (ICCCNT), July 2018, pp 942–945
15. Lingayat A, Badre RR, Gupta AK (2018) Integration of linux containers in openstack: an introspection. *Indonesian J. Electr. Eng. Comput. Sci.* 12(3)
16. Lingayat A, Naik V, Singh A, Wankhade S, Dhobale N (2018) User interface for managing openstack magnum service using openstack horizon. In: Two days 2nd national level conference on emerging trends in computer engineering and technology (NCETCET18), Jan 2018, pp 386–390
17. Tosatto A, Ruiu P, Attanasio A (2015) Container-based orchestration in cloud: state of the art and challenges. In: 2015 ninth international conference on complex, intelligent, and software intensive systems. IEEE, pp 70–75
18. Cacciatore K, Czarkowski P, Dake S, Garbutt J, Hemphill B, Jainschigg J, Moruga A, Otto A, Peters C, Whitaker BE (2015) Exploring opportunities: containers and openstack. OpenStack white paper, vol 19

19. Menage P (2018) “Cgroups” [Online]. Available: <https://www.kernel.org/doc/Documentation/cgroup-v1/cgroups.txt>
20. Felter W, Ferreira A, Rajamony R, Rubio J (2015) An updated performance comparison of virtual machines and linux containers. In: 2015 IEEE international symposium on performance analysis of systems and software (ISPASS). IEEE

# A Review of Metaheuristic Techniques for Solving University Course Timetabling Problem



Manpreet Kaur and Sanjay Saini

**Abstract** Educational timetable generation is one of the major administrative requirements in schools and universities. University course timetabling problem falls in the category of NP-hard problems having various constraints, objectives, and limited resources. Generating an optimized timetable is challenging and time-consuming process. The objective here is to present a concise review of some recent techniques that researchers have tried to resolve university course timetabling problem having single/multiple objectives.

**Keywords** University course timetabling problem (UCTTP) · Multi-objective optimization · Educational timetabling problems · Metaheuristic techniques

## 1 Introduction

Many of the optimization problems in the real world have multiple objectives. In most of the cases, these objectives are conflicting which prevents simultaneous optimization. That means the objectives are defined in incomparable units, and they present some degree of conflict among them. There are numerous methods followed in the literature to solve multi-objective optimization problems [1]. There are two general methodologies for multiple objective optimizations:

1. Problem scalarization (preference-based classical approach)
2. Pareto optimal solution (ideal approach).

One of the scheduling problems which are catching the attention of researches over few decades is timetabling problem. There are various types of timetabling problems that have emerged in recent times, for example, university and school timetabling, public transportation timetabling, tournament scheduling, and television programs

---

M. Kaur (✉) · S. Saini

Department of Physics and Computer Science, Dayalbagh Educational Institute, Agra, India  
e-mail: [new.meenu@gmail.com](mailto:new.meenu@gmail.com)

S. Saini  
e-mail: [sanjay.s.saini@gmail.com](mailto:sanjay.s.saini@gmail.com)

scheduling. Depending on the problem instance, the constraints and objectives of the problem can vary. In one case, the objective may be to minimize the total time period for scheduling tasks, yet in others, the objective can just be finding a feasible solution. Multi-objective timetabling problem can have two or more contradictory objectives. The optimization of one objective may lead to minimization of the number of clashes; the other objective may tend to decrease the length of student/teacher's time span.

## 2 Educational Timetabling Problems

In wide variety of institutions, educational timetabling is one of the major administrative requirements. Timetabling problem can be defined as “Allocation of given resources to objects being placed in time slots in such a way as to satisfy, as nearly as possible, a specified set of constraints” [2, 3].

Timetabling problems have lots of constraints that can be categorized into following two types [4]:

- Hard constraints
- Soft constraints.

Hard constraints are those that are strictly imposed. Violation of any one of the hard constraints will result in an infeasible timetable. Few examples of hard constraints are: At any given time, resource either student or teacher can be expected to be in one place only; The room capacity has to be greater than or at least equal to total number of students attending the event; only one course has to be allotted to any room in the same timeslot; For every timeslot, enough resources must be there.

Soft constraints are desirable to be imposed but not mandatory. Examples of such soft constraints are: A course should be allotted to particular time slot; a course may be required to schedule subsequent to the other; two exams should not be scheduled for one student on the same day, etc.

Educational timetabling problems are mainly categorized in three main types:

- Course timetabling
- Examination timetabling
- School timetabling.

However, all these educational timetabling problems seem to be similar but a technique to resolve one type of timetabling problem may not resolve another type giving the same level of success. No solution can be said better than the other and each solution corresponds to the trade-offs between different objectives [5].

The importance of this problem is primarily because of the complexity of generating a feasible timetable that satisfies the preferences of the administration, the teachers, and the students. In few of cases, a single feasible solution may not be

found. It is almost impossible to prepare a universal representation which is appropriate for every case, since all institutes have their own dimensions, objectives, and set of constraints.

### 3 Techniques to Solve Timetabling Problems

Main techniques used for solving the university timetabling problem are discussed below [6]:

#### 3.1 Operations Research Based Techniques

These techniques include graph coloring, integer linear programming, and constraint satisfaction(s) programming.

Graph coloring theory is capable of generating conflict-free timetables but graph coloring method does not have performance efficiency. Hybridization with any other metaheuristic approach can be used to increase the performance. Welsh and Powell solved timetabling problem using graph coloring problem in 1967 [7]. Asham, Soliman, and Ramadan proposed a hybrid approach using genetic coloring for reducing the cost of finding the chromatic number which is the minimum number of required colors to color the graph [8].

Integer/linear programming is a mathematical method. It is implemented on basis of the type and structure of the institute. Hybridization with other approaches is not done in this technique, though constructive heuristics can be used with this method for analyzing the constraints. An integer programming technique has been used by Bakir and Aksop that resulted in the lessening of disappointment of teachers and students while simultaneously applying a specific set of rules [9].

The constraint satisfaction programming is a system based on computing. Deris, Omatu, and Ohta resolved a timetabling problem with help of constraint-based reasoning technique in an object-oriented approach. Zhang and Lau applied ILOC software to apply the CSP approach to generate timetables in university [10].

#### 3.2 Metaheuristic Techniques

These techniques include case base reasoning approach (CBR), single solution-based techniques, and population-based techniques.

Population-based methods comprise the following algorithms: evolutionary and genetic algorithms (GAs), ant colony optimization methods (ACO), particle swarm optimization technique (PSO), intelligent water drops (IWD), etc.

Charles Darwin proposed evolutionary algorithms (EAs) in 1859. Evolutionary algorithms are based on the simplified biological model of evolution and natural selection. EAs work on a group of possible solutions. Genetic algorithms are a type of EAs. Recently in 2016, Shaikh et al. presented genetic algorithm with metaheuristic approach for generating timetable [11]. Abdelhalim used genetic algorithm to produce good quality feasible timetables. They used some heuristics that generate an initial population [12].

Ant colony optimization is one of the metaheuristic algorithms with built-in capability of optimization that finds an optimal solution over a discrete search space. They use probabilistic rules to find the shortest route between two points. In the early 1990s, Macro Dorigo et al. introduced ACO by observing the foraging behavior of ant colonies. Socha et al. used max-min ant system for the generation of university course timetable by constructing an optimal path. Every path could produce a constructive graph to allocate courses to timeslots based on the amount of pheromone [13]. Mayer, Nothegger, Chwatal, and Raidl applied ACO algorithm for the UCTTP on ITC-2007 dataset. This algorithm has good performance with faster run time [14]. Ayob and Jaradat proposed two types of hybrid ant colony to solve the UCTTP [15]. These include hybridization of ant colony with simulated annealing and secondly with tabu search. Curriculum-based UCTTP has been generated by Patrick, Godswill using greedy ant colony optimization technique [16]. Mazlan et al. implemented and tested ACO algorithm approach in Web-based computer system. They find this approach reliable to be applied on UCTTP [17].

After taking inspiration from the study of the behavior of bird flocking, biologist Frank Heppner, James Kennedy, and Russell Eberhart developed PSO in 1995. PSO approach can be used to solve the problems whose results can be depicted as a point in an n-dimensional solution space [18]. Chen and shih used particle swarm optimization (PSO) on UCTTP. Montero et al. applied PSO with local search procedure to find high-quality solutions [19]. Fong et al. presented a hybrid swarm-based technique using a variation of artificial bee colony algorithm.

Single solution-based approaches do not use initial population to find solution to problems. These approaches use a single solution to analyze the optimization which is chosen on the basis of some criteria. A few single solution algorithms are tabu search algorithm (TS), Great Deluge Algorithm (GD), variable neighborhood search (VNS), simulated annealing (SA), etc.

Hertz presented tabu search algorithm for generating solutions to large-scale timetabling problems for decreasing the conflict occurrence resulting due to events going on at the same time but involving common students or staff, or entailing the same room [20]. The impact of neighboring structures on tabu search algorithm has been presented by Aladag, Hocaoglu, and Basaran to solve the UCTTP. Impact of simple search and swap moves is tested on neighboring structures-based tabu search [21].

Aycan and Ayav presented simulated annealing approach for solving the UCTTP [22]. A comparison of the effectiveness of different neighboring search algorithms is performed by them.

### ***3.3 Intelligent Novel Approaches***

Intelligent novel approaches incorporate following methods: hybrid approaches; fuzzy theory-based approaches; artificial intelligence-based approaches; and clustering algorithms. In terms of efficiency and performance, hybrid approaches have produced attractive results while solving NP-complete problems [23].

A hybridization of algorithms used sequential heuristic with simulated annealing to generate solution to the UCTTP using ITC-2002 dataset by Kostuch. Abdullah et al. presented a hybrid evolutionary approach to solve the UCTTP where the combinations of local search algorithm and evolutionary approaches give improved solutions. Shahvali Kohshori and Saniee Abadeh proposed a hybrid genetic algorithm to solve UCTTP which is based on FGARI, FGASA, and FGATS genetic algorithms. Badoni and Gupta presented a hybrid approach to solve UCTTP which unites the qualities of genetic algorithm with iterated local search algorithm [24].

Fuzzy logic is among the main principles of computational intelligence proposed by Lotfi Zadeh in 1965. Most important feature of fuzzy logic is that it can face incompleteness. Fuzzy logic works with partial truth values lying between exact zero and exact one. To solve a resource allocation problem with multiple objectives, a hybrid fuzzy evolutionary algorithm was proposed by Rachmawati and Srinivasan [25]. Shahvali Kohshori et al. came up with a fuzzy genetic algorithm in combination with local search for giving solution to UCTTP [26].

### ***3.4 Distributed Multi-agent Based Approaches***

This system comprises of multiple agents that are intelligent enough and they interact with each other. Every agent can observe and obtain input from environment through sensors and then action is performed using a driver. According to the application of different agents, they are categorized into different classes, such as autonomous, intelligent, actionable, proactive, learner, mobile, and cooperative/communicative agents. Gaspero, Missaro, and Schaerf implemented distributed multi-agent architecture to generate course timetables which consist of a set of courses in predetermined timeslots in a circulating week [27]. The study of UCTTP as distributed timetabling problem was carried out by Xiang and Zhang [28]. Nandhini and Kanmani implemented course timetable based on multi-agent systems where hill climbing algorithm was used with sharpest upward. Yang and Paranjape showed in their study that implementing multi-agent system needs development of an intelligent decision-making system [29].

## 4 Conclusion

So far, many methods have been proposed, discussed, and evaluated by different researchers for the educational timetabling problem, although there is still work to be done in the future. There is no method that can be used to generate feasible solution for all instances of educational timetabling problems. Educational timetabling problems having conflicting multiple objectives such as minimizing teacher's time span and minimizing student's time span can be worked upon. Hybrid methods have been found to be the most effective to solve timetabling problems. A few of the potential things that can be taken up to create educational timetables are as follows:

- Using the hybridization of swarm intelligence-based techniques to obtain feasible solutions to the educational timetabling problems in order to underpin the next generation of general and adaptive timetabling algorithms.
- Designing parallel algorithms to create the educational timetables with the available methods to increase their efficiencies.
- The integration of methods, such as integer linear programming and constraint programming with metaheuristics for solving large and highly constrained timetabling problems.

## References

1. Deb K (2005) Multiobjective optimization using evolutionary algorithms. Wiley, Chichester
2. Burke EK, Petrovic S (2002) Recent research directions in automated timetabling. *Eur J Oper Res* 140(2):266–280
3. Carter MW (2000) A comprehensive course timetabling and student scheduling system at the University of Waterloo. In: Practice and theory of automated timetabling III, pp 64–84
4. Cruz-Chávez MA et al (2016) Solving a real constraint satisfaction model for the university course timetabling problem: a case study. *Math Probl Eng*. <http://dx.doi.org/10.1155/2016/7194864>
5. Pillay N (2016) A review of hyper-heuristics for educational timetabling. *Ann Oper Res* 239:3–38. <https://doi.org/10.1007/s10479-014-1688-1>
6. Babaei H et al (2015) A survey of approaches for university course timetabling problem. *Comput Ind Eng* 86:43–59
7. Welsh DJA, Powell MB (1967) An upper bound for the chromatic number of a graph and its application to timetabling problems. *Comput J* 10:85–86
8. Asham GM, Soliman MM, Ramadan AR (2011) Trans genetic coloring approach for timetabling problem. *Artif Intell Tech Novel Approaches Pract Appl IJCA* 17–25
9. Bakir MA, Aksop C (2008) A 0–1 integer programming approach to a university timetabling problem. *Hacettepe J Math Stat* 37(1):41–55
10. Zhang L, Lau SK (2005) Constructing university timetable using constraint satisfaction programming approach. In: IEEE Proceedings of the 2005 international conference on computational intelligence for modeling, control and automation, and international conference on intelligent agents, web technologies and internet commerce (CIMCA-IAWTIC'05), vol 2, pp 55–60
11. Shaikh A et al (2016) Genetic algorithm with meta-heuristic approach for generating timetable. *Int J Adv Res* 4(2):300–304

12. Abdelhalim EA et al (2016) A utilization-based genetic algorithm for solving the university timetabling problem (UGA). *Alexandria Eng J* 55:1395–1409
13. Socha K et al (2002) A max-min ant system for the university course timetabling problem. In: Proceedings of the 3rd international workshop on ant algorithms (ANTS 2002), Lecturer notes in computer science, 2463. Springer, Berlin, pp 1–13
14. Mayer A, Nothegger C, Chwatal A, Raidl G (2008) Solving the post enrolment course timetabling problem by ant colony optimization. In: Proceedings of the 7th international conference on the practice and theory of automated timetabling
15. Ayob M, Jaradat G (2009) Hybrid ant colony systems for course timetabling problems. In: 2nd conference on data mining and optimization, pp 120–126
16. Patrick K, Godswill Z (2016) Greedy ants colony optimization strategy for solving the curriculum based university course timetabling problem. *Br J Math Comput Sci* 14(2):1–10
17. Mazlan M et al (2018) Ant colony optimisation for solving university course timetabling problems. *Int J Eng Technol* 7(2,15):139–141
18. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of the IEEE international conference on neural networks, Perth, pp 1942–1945
19. Montero E et al (2011) A PSO algorithm to solve a real course + exam timetabling problem. In: International conference on swarm intelligence, 24, pp 1–8
20. Hertz A (1991) Tabu search for large scale timetabling problems. *Eur J Oper Res* 54:39–47
21. Aladag CH, Hocaoglu G, Basaran MA (2009) The effect of neighborhood structures on tabu search algorithm in solving course timetabling problem. *Expert Syst Appl* 36:12349–12356
22. Aycan E, Ayav T (2009) Solving the course scheduling problem using simulated annealing. In: Advance computing conference, IACC. IEEE International
23. Henry Obit J (2010) Developing novel meta-heuristic, hyper-heuristic and cooperative search for course timetabling problems. Ph.D. thesis, School of Computer Science, University of Nottingham
24. Badoni RP, Gupta DK (2015) A hybrid algorithm for university course timetabling problem. *Innovative Syst Des Eng* 6(2):60–66
25. Rachmawati L, Srinivasan D (2005) A hybrid fuzzy evolutionary algorithm for a multi-objective resource allocation problem. In: IEEE proceedings of the fifth international conference on hybrid intelligent system
26. Kohshori MS, Abadeh MS (2012) Hybrid genetic algorithms for university course timetabling. *Int J Comput Sci Issues* 9(2):2
27. Di Gaspero L, Mizzaro S, Schaerf A (2004) A multi-agent architecture for distributed course timetabling. In: Proceedings of the 5th international conference on the practice and theory of automated timetabling (PATAT '04), pp 471–447
28. Xiang Y, Zhang W (2008) Distributed university timetabling with multiply sectioned constraint networks. In: Proceedings of the twenty-first international FLAIRS conference
29. Yang Y, Paranjape R (2011) A multi-agent system for course timetabling. *Intell Decis Technol Comput Sci Artif Intell* 5(2):113–131

# Using Social Media Analytics to Predict Social Media Engagement Outcome for Fortune CEOs



Hitesha Yadav, Arpan K. Kar, and Smita Kashiramka

**Abstract** Social media has been broadly adopted by corporates for communicating and networking. Social media platforms help firms in building a relationship with the outside world. The study highlights how the adoption of social media is changing the way firms connect with their stakeholders. It highlights how CEOs' social media engagement on behalf of the firm contributes to building up firms' reputation amongst consumers. Influencer CEOs from top Fortune 500 companies have been used for this purpose. Social media analytics has been used in this study to get relevant insights from Twitter using methods of content analytics. Mining of user-generated content and further analysis of its impact on engagement outcome of CEOs has been done using methods such as web-scraping, topic modelling, bag-of-words, and dictionary methods. The study analyses CEOs' engagement with tweets specific to a firm's orientation towards adopting sustainable development goals (SDG). The study reveals how SDG-related messages on the Twitter timeline of CEOs have high user engagement and helps build customer reputation.

**Keywords** Social media · Twitter analytics · Social media engagement · Sustainability · Corporate reputation · Data mining · Big data analytics

## 1 Introduction

The top priority of a CEO is to build healthy connections between the firm and their stakeholders. Social media adoption has now become a new way of communication, growing at a rapid and dynamic pace [1]. It is two-way communication in a distributed

---

H. Yadav (✉) · A. K. Kar · S. Kashiramka

Department of Management Studies, Indian Institute of Technology, Delhi, India

e-mail: [hitesha1902@gmail.com](mailto:hitesha1902@gmail.com)

A. K. Kar

e-mail: [arpan.kumar.kar@gmail.com](mailto:arpan.kumar.kar@gmail.com)

S. Kashiramka

e-mail: [smitakashiramka@gmail.com](mailto:smitakashiramka@gmail.com)

environment for generating content, circulation and communication amongst online groups. According to the past literature the customers aren't enthusiastic to engage with a firm's social media [2]. CEOs' strategic adoption of the social media platforms has drastically transformed the way consumers perceive about the firm [3, 4]. CEOs interacting with stakeholders on behalf of the firm over social media have helped in building a healthy relationship, overcoming the communication barrier [5, 6]. Further, CEOs have traditionally held huge influential power over different stakeholders of the firm like employees, government and other citizens.

The United Nations (UN) proposed the sustainable development goals (SDGs), providing the guidelines to the member states towards sustainable and equitable society that will lead to long-term implications on the business organizations. Organizations can adopt these SDGs as a guideline for investing in sustainable development while not compromising their own business interests [7, 8]. SDGs focus is for improved natural environment, developing effective infrastructure and well-being of individual as well as community. Based on their gross revenue, a list of top 500 companies is released every year by Fortune magazine as Fortune companies [9]. Therefore, for our study we chose to explore how the adoption of Twitter by CEOs of Fortune 500 companies for their strategic expression contributes to their corporate reputation. 53 CEOs, who were active and hence influential, were used in this analysis by doing a topic modelling of timeline Tweet extraction using Latent Dirichlet Algorithm (LDA). Subsequently Automated Content Analysis (ACA) has been used to model the topics derived out of LDA. There are several social media platforms available in the market but for our study we took Twitter. The reasons being, twitter data is easily accessible through Twitter application programming interface (API). The user timeline could be downloaded, and analysis could be performed as per the requirement of the study. Also, Twitter is the fastest-growing platform with increasing number of users and vast geographical reach.

Thus, this paper tries to capture the strategic expression of Fortune CEOs and then tries to determine if adoption of SDG on their social media engagement leads to reputation through posts on Twitter using LDA and ACA. The rest of the paper consists of five sections: conceptual framework and research question for the study, methodology describing the research process undertaken to conduct this study, findings and analysis, followed by discussions and conclusion.

## 2 Conceptual Framework and Research Question

Adoption of social media networks explains the adoption phenomenon as an interactive process between organisations and its environment. Technology can be effortlessly used to develop public relations and spread the message on what the firm is focussing on to the stakeholders. For social media to be successfully exploited for their utility, firms need to design experiences to the customers that deliver tangible value in return for their time, attention, endorsement and data [4].



**Fig. 1** Conceptual framework of the study

Firm's orientation towards adopting social development goals (SDG), practicing which reduces the negative environmental impacts, simulating new business opportunities and facilitates policy integration across sectors [10]. Methods of social media analytics (SMA) such as NLP, opinion mining, scraping, sentiment analysis, text analytics are used in collecting, analysing, summarizing and visualising the social media data for a specific requirement [11, 12].

Figure 1 represents the conceptual model for this study, trying to investigate whether tweets related to different SDGs posted by Fortune CEOs have any impact on their engagement within the social media community by attracting more shares, likes and having more social media members following them. The following research question has been formulated to support the above investigation.

**RQ.** How does discussions surrounding sustainable development goals (SDG) by a firm's CEO on social media lead to their social media reputation?

The next section tries to explain how different methods of social media analytics and machine learning have been applied to CEOs social media data to conduct the study.

### 3 Methodology

The empirical research methods such as surveys, interviews, discussions and qualitative case studies cannot be used for conducting this study. Since the respondents of the study are the CEOs of Fortune companies, access to all of them is very difficult and there is a high probability they will not respond to the survey or in lines of socially desirable responses.

In the literature, it is mentioned that most of the surveys conducted by the researchers have been consistently neglecting biases related to social desirability response as well as self-presentation [13, 14]. Further, the theme being a sensitive one, CEOs are more likely to provide a response, which suffers from high social desirability biases and self-presentation biases. Hence, this study makes use of the timeline of CEOs on Twitter because CEOs messages would indicate firm's orientation towards the adoption of SDGs on a personal front. This will eventually represent the way firms would operate and perform under their leadership. The Twitter timeline of the firms were not considered for our study as many of them might hold multiple accounts and managed professionally by agencies, which may lead to biases.

The methodology section has been divided into two subsections. The first subsection illustrates the various data collection methods adopted and the second subsection elaborates on the categorization of tweets to seventeen different SDGs using methods of content analysis.

### 3.1 Data Collection

Fortune 500 CEOs were searched on Google Search box as well as Twitter, by putting the CEO name along with the company name. The Google Search results were screened for the presence of Twitter handle of individual Fortune CEOs. The timeline of all the CEOs present on Twitter was then scrapped for conducting further analysis.

It was found that out of the Fortune 500 only 73 CEOs were present on Twitter. Amongst these 73 CEOs, only 58 were active on Twitter, whereas the non-active CEOs were then removed for further analysis.

Twitter API had been used for timeline extraction of Fortune CEO's. The tweets extracted contain the following variables: identification number of each tweet (id), creation date of the tweet/post (created\_at), tweet content/message (text), number of times tweets liked by the users (favourite\_count), tweet sharing count (retweet\_count), hashtags present in the tweet along with the indices (hashtags), other twitter users mentioned in the tweet (user\_mentions). Bloomberg terminal was used to collect other firm related details such as Fortune rank of the firm, employee number, sector, industry.

### 3.2 Data Analysis

Methods of social media analytics in combination with conventional data mining and analysis approaches were used so as to derive the insights from the available data. Majorly, social media analytics [15, 16], big data analytics [17–19] and machine learning techniques [8, 15] were used as a methodology in this study.

Content analysis method has been used to recognize the CEO's opinion on various SDGs on Twitter. Content analysis technique belongs to the domain of natural language processing (NLP) and text mining in computer science area. Based on inherent semantics of the content, it helps in converting the qualitative data into quantitative data [12, 14, 20]. Analysing the quantitative data using statistical approaches, then followed this.

It is very difficult to handle the large data volume manually; thus for analysing the large collection of CEO tweets, automatic content analysis method was adopted. In computer science, automated content analysis methods are used to handle large volume of data, which are optimized to classify individual documents [21–23]. Automatic content analysis tracks linguistic patterns across enormous textual data and also reduces the cost of analysis [24, 25].

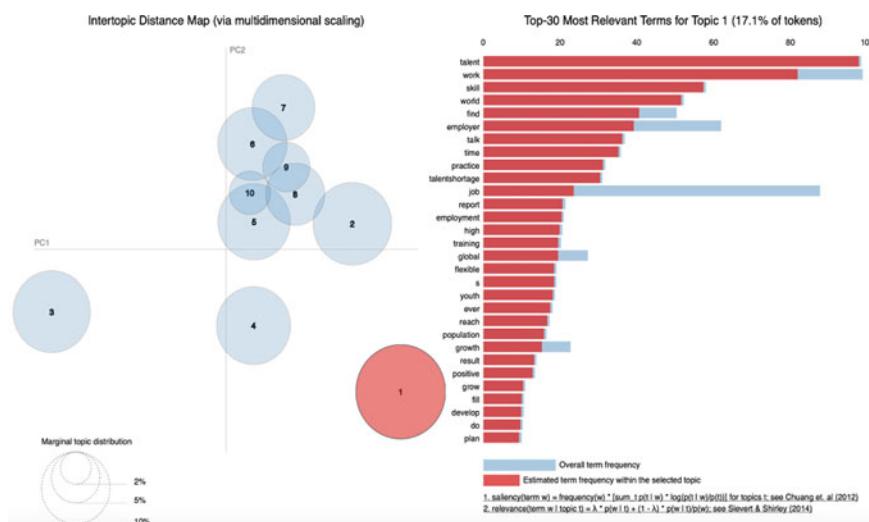
Tweets extracted from the timeline are unstructured. The content consists of text, hyperlinks, images, hashtags, etc. To extract useful information from this unstructured text for further analysis, unwanted text such as URLs, punctuations, stop words were removed in pre-processing stage. After the tweets were cleaned, all the words

were then converted to lowercase format. Topic modelling algorithm and bag-of-words methods were adopted to obtain the most dominant and frequently used keywords in each CEO's tweet posts [22, 26]. For each CEO, ten topics with ten keywords were extracted to get the most prevalent words in the text.

Dictionary method is a data-driven technique that classifies the documents based on the keywords that appear in the text. With a dictionary that is well-defined, frequency count of the top keywords appearing in each topic provides a dynamic and reliable analysis of the text. This method is also referred to as frequency or categorical analysis [27]. A dictionary containing rules for categorization of the topics was prepared in the form of a codebook with reference to the guidelines provided by United Nations (UN) on sustainability. The topics derived from ACA were then mapped with the keywords in the dictionary for SDG scoring. A sample of inter-topic mapping for relevant topics in a CEO's tweets is highlighted in Fig. 2.

A score out of 5 (using Likert scale of measurement; 1—low presence of the keywords for topics in the codebooks, 5—high presence of the keywords for topics in the codebooks) had been allotted depending on the weightage of SDG keywords mentioned in Annexure I. Even if the tweets contained words for more than one SDG, the score was allotted to each. An average of all the scores was then taken to derive the final SDG score corresponding to each CEO's tweets.

Three variables were considered to compute the CEOs reputation on Twitter, Follower Count (FC), Likes Count (LC) and Retweet Count (RC) were obtained by averaging out the number of likes and tweets shared by others of the tweets posted by each CEO on their twitter profile and followed by some statistical methods to check for hypotheses.



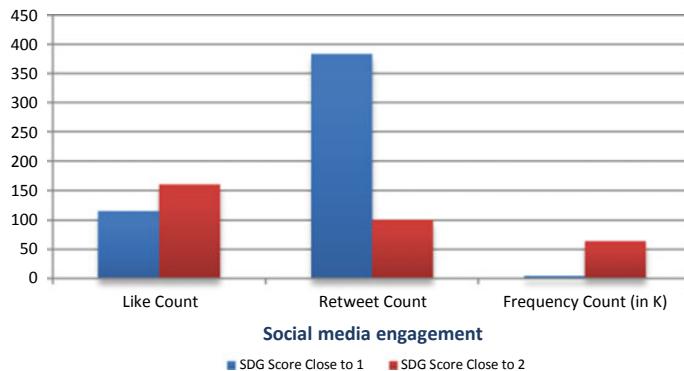
**Fig. 2** Inter-topic mapping for most relevant topics in CEO tweets

## 4 Findings and Analysis

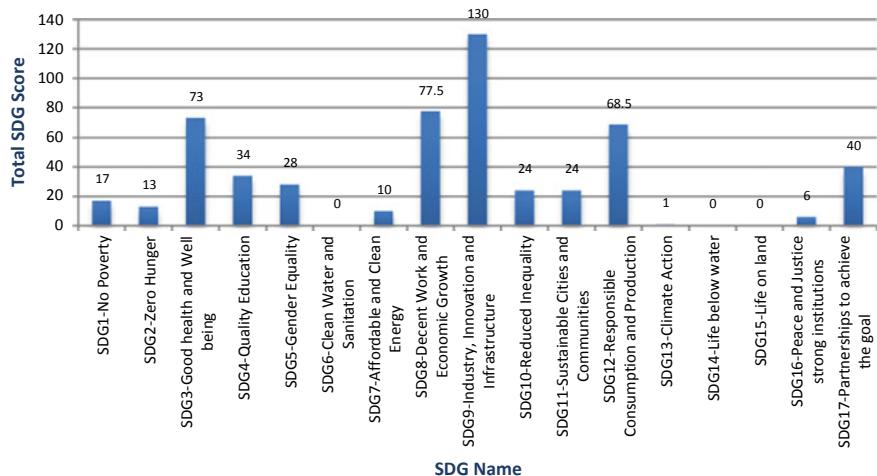
It was observed that only 14.6% amongst 500 Fortune CEOs are present on Twitter. 83.56% were found to be active on Twitter amongst these Fortune CEOs.

Figure 3 presents the comparison of Follower Count (FC), Likes Count (LC) and Retweet Count (RC) for the CEOs with SDG score close to 1 and 5. FC and LC are more for CEOs with SDG score close to 5, whereas RC is more for CEOs with SDG score close to 1.

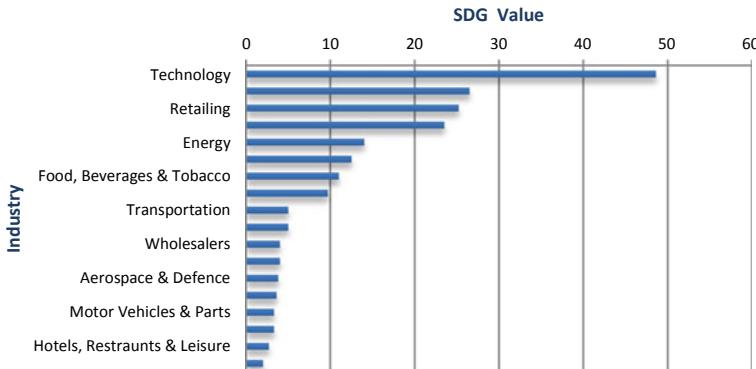
Figure 4 presents the CEOs tweets surrounding individual SDGs. CEOs had tweeted the most on industry, innovation and infrastructure (SDG9) followed by



**Fig. 3** Engagement of CEO tweets having SDG score close to 1 and 5



**Fig. 4** Most prevalent SDGs in CEOs tweets



**Fig. 5** Focus of SDGs in CEO tweets sectorwise

good health and well-being (SDG3), decent work and economic growth (SDG8) and responsible consumption and production (SDG12).

As highlighted in Fig. 5, technology, financials, retailing and business services are the sectors in which CEOs tweet more about SDGs. Only some of the sectors are inclined towards adopting SDGs, and there are few selective SDGs that are most popular amongst the CEOs messages on social media.

For Fortune CEOs regression, analysis was done according to which there is a significant relationship between the CEO's adoption of SDGs in their tweets in terms of likes and retweets with  $p < 0.5$ . Hence, there is a statistically significant implication on liking and re-sharing of the tweets posted by Fortune CEOs in SDG context.

All the above findings are clear indications that the social media engagement of CEOs related to SDG tweets shows a positive outcome, thus supporting RQ1.

## 5 Conclusion

By using the data mining and social media big data analytics methods such as content analysis, topic modelling, ACA, bag-of-words, dictionary method, this study tries to investigate the relationship between CEOs strategic expression and engagement on social media using Twitter handle. The exploration in the study reveals high engagement of users in terms of likes and follower count when Fortune CEOs are extensively posting on SDGs as compared to other tweets. Also, CEOs from technology and financial sectors are posting more about SDGs compare to household products or hotels, restaurants and leisure sectors. This study can contribute significantly for strategizing the adoption and effective use of social media by CEOs to develop higher engagement and connect with the consumers, in turn building high reputation of the firm.

CEO of an organization holds an important role in shaping up of their companies' social reputation and has many agendas on their priority list defined in their mission statements [13, 28]. In future, apart from SDGs there are more possible topics of concern, which can be taken up to extend this study further. In future, researchers may explore more areas that are on CEOs priority list such as circular economy (CE); corporate governance; environment, social and governance goals (ESG).

## References

1. Grover P, Kar AK, Ilavarasan PV (2019) Impact of corporate social responsibility on reputation—insights from tweets on sustainable development goals by CEOs. *Int J Inf Manage* 48:39–52
2. Hwang S (2012) The strategic use of Twitter to manage personal public relations. *Public Relat Rev* 38(1):159–161
3. Grover P, Kar AK (2018) User engagement for mobile payment service providers—introducing the social media engagement model. *J Retail Consum Serv* 53
4. Hong L, Davison B (2010) Empirical study of topic modelling in Twitter. In: SOMA'10 proceedings of the first workshop on social media analytics, pp 80–88
5. Capriotti P, Ruesja L (2018) How CEOs use Twitter: a comparative analysis of global and Latin American companies. *Int J Inf Manage* 39:242–248
6. Zhou M, Lei L, Wang J, Fan W, Wang AG (2014) Social media adoption and corporate disclosure. *J Inf Syst* 29(2):23–50
7. Korhonen J, Honkasalo A, Seppälä J (2018) Circular economy: the concept and its limitations. *Ecol Econ* 143:37–46
8. Chakravorti B (2017) How companies can champion sustainable development. *Harv Bus Rev*. <https://hbr.org/2017/03/how-companies-can-champion-sustainable-development>. Last accessed 2019/01/3
9. Fortune 500 list. <https://www.someka.net/excel-template/fortune-1000-excel-list/>. Last accessed 2019/01/15
10. Batrinca B, Treleaven P (2015) Social media analytics: a survey of techniques, tools and platforms. *AI Soc* 30:89–116
11. Stieglitz S, Dang-Xuan L, Bruns A, Neuberger C (2014) Social media analytics. *Bus Inf Syst Eng* 6(2):89–96
12. Baird C, Parasnis G (2011) From social media to social customer relationship management. *Strategy Leadersh* 39(5):30–37
13. King MF, Bruner GC (2000) Social desirability bias—a neglected aspect of validity testing. *Psychol Mark* 17(2):79–103
14. Arkin RM, Gabrenya WK Jr, Appelman AS, Cochran ST (1979) Self-presentation, self-monitoring, and the self-serving bias in causal attribution. *Pers Soc Psychol Bull* 5(1):73–76
15. Rathore A, Kar AK, Ilavarasan PV (2017) Social media analytics—literature review and directions for future research. *Decis Anal* 14(4):229–249
16. Kar AK (2015) Integrating websites with social media—an approach for group decision support. *J Decis Syst* 24(3):339–353
17. Grover P, Kar AK (2017) Big data analytics—a review on theoretical contributions and tools used in literature. *Global J Flex Syst Manage* 18(3):203–229
18. Joseph N, Kar AK, Ilavarasan V, Ganesh S (2017) Review of discussions on Internet of Things (IoT): insights from Twitter analytics. *J Global Inf Manage* 25(2):37–60
19. Schwartz H, Ungar LH (2015) Data-driven content analysis of social media: a systematic overview of automated methods. *Am Acad Polit Soc Sci* 659:78–94

20. Elo S, Kaariainen M, Kanste O, Polkki T, Utriainen K, Kyngas H (2014) Qualitative content analysis—a focus on trustworthiness. *SAGE Open* 4(1):2158244014522633
21. Nunez-Mir GC, Iannone BV III, Pijanowski BC, Kong N, Fei S (2016) Automated content analysis—addressing the big literature challenges in ecology and evolution. *Meth Ecol Evol* 7(11):1262–1272
22. Elo S, Kyngäs H (2008) The qualitative content analysis process. *J Adv Nurs* 62(1):107–115
23. Wallach H (2006) Topic modelling-beyond bag-of-words. In: ICML'06 proceedings of the 23rd international conference on machine learning, pp 977–984
24. Verschoor CC (2003) Corporate responsibility—high priority for CEOs. *Strateg Finance*
25. Hopkins D, King G (2009) A method of automated nonparametric content analysis for social science. *Am J Polit Sci* 54:229–247
26. Krippendorff K (2004) Content analysis—an introduction to its methodology. Sage, Beverly Hills, CA
27. Le Blanc D (2015) Towards integration at last? The sustainable development goals as a network of targets. *Sustain Dev* 23:176–187
28. Kassarjian HH (1997) Content analysis in consumer research. *J Consum Res* 4(1):8–18

# Smart Heart Attack Forewarning Model Using MapReduce Programming Paradigm



Arushi Jain, Vishal Bhatnagar, and Annavarapu Chandra Sekhara Rao

**Abstract** The information and communication technology (ICT)-related exponential growth has increased the demand for big data analytics (BDA). BDA involves the handling of a gigantic data for storage and investigation. The evolving field of BDA owns many challenges in various fields including drug delivery, healthcare, surveillance, weather forecasting, etc. In comparison with other industries, the need for big data in healthcare experiences more attention in present days. Initially, the data collected from remote healthcare services vary based on value, variety, velocity, veracity, and volume since the collection occurs at different locations using various devices. In research and development, there is an urge for an algorithm in risk prediction of heart attack. One of the major diseases related to mortality is cardiovascular disease (CVD). Further, an approach is introduced, and this approach has improved performance in terms of accuracy of 99%. However, in future works, it is recommended to focus on various other nature-inspired algorithms for diseases such as thyroid, diabetes, and so on.

**Keywords** Big data analytics · Hadoop · MapReduce · Cardiovascular disease · Heart attack

---

A. Jain (✉) · A. C. S. Rao  
Indian Institute of Technology, Dhanbad, India  
e-mail: [arushijain1391@gmail.com](mailto:arushijain1391@gmail.com)

A. C. S. Rao  
e-mail: [chandra.annavarapu@gmail.com](mailto:chandra.annavarapu@gmail.com)

V. Bhatnagar  
Ambedkar Institute of Advanced Communication Technologies and Research, Geeta Colony,  
New Delhi 110031, India  
e-mail: [vishalbhatnagar@yahoo.com](mailto:vishalbhatnagar@yahoo.com)

## 1 Introduction

Traditionally, medical data was in form of hospital records or national statistics registries. With advancement in computational science and increased dependence on digital data, mass datasets of medical data or medical big data is generated. Generally, big data is defined as an emerging use of rapidly collected, complex data requiring storage size of terabytes, petabytes, or zettabytes. The unique properties of big data are defined by four dimensions: volume, velocity, variety, and veracity [1]. Medical big data analytics is useful in addressing the growing crisis of the chronic disease. It is a revolutionary technology which uses machine learning algorithm models to collect, generate, and organize huge volumes of multi-structured information retrieved from the hospital and clinical records, national and international data repositories, medical transcripts or from personal body sensors and wearables [2]. Big data processing is achieved through various techniques including cloud environment, genotyping, artificial neural network, K-nearest neighbor, particle swarm optimization technique, HADOOP, naive Bayes, C4.5 algorithm, decision tree, IB1 classifier, radial basis function (RBF), etc. Incidentally, in a cloud environment, entire resources are set at a federal situation using ICT-based infrastructure thereby delivering remote healthcare services [3]. Such devices may be found ubiquitously in diverse locations. The cloud-based approaches are used in various splits of healthcare field including genomics, pharmaceuticals, medicine, research and development, etc. However, in research and development, there is an urge for an algorithm in risk prediction of heart (cardiovascular) diseases. In particular, cardiovascular disease (CVD) is one of the major diseases related to mortality. CVD consists of numerous complications, with diverse systems and symptoms including arrhythmia, cardiac arrest, congestive heart failure, hemorrhagic stroke, heart valve problem, ischemic stroke, and myocardial infarction (MI) or heart attack [4]. CVD is unique amid the list of ailments ranked for consideration by National Health Priority Area (NHPA) initiative. Health safeguard expenditures for CVD are mounting because of the expanding frequency of diabetes, metabolic syndrome, obesity, and stress [5].

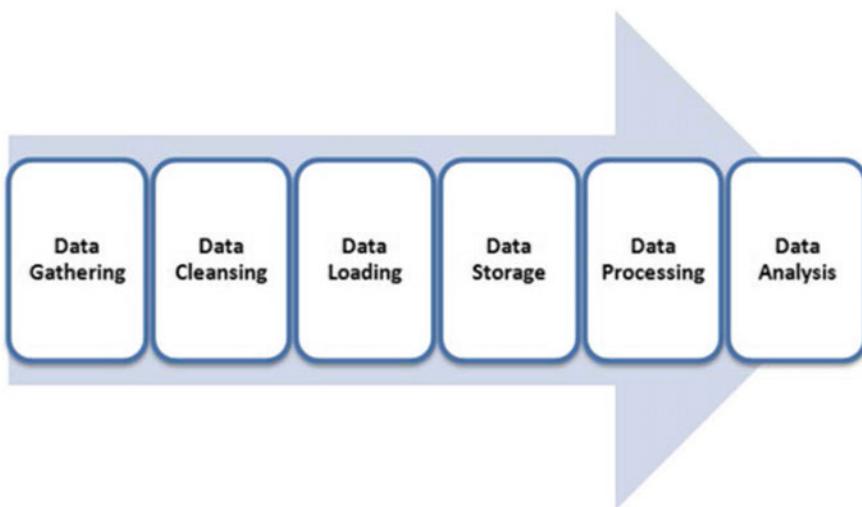
## 2 Literature Review

Chen et al. [6] recommended a new multimodal disease risk prediction (CNN-MDRP) algorithm based on convolutional neural network. In relation to numerous distinct prediction algorithms, the forecasting precision of algorithm attains at 94.8% with a convergence speed that is quicker than that of the CNN-based unimodal disease algorithm. Chen et al. [7] proposed an integrated system for concurrent feature selection and fuzzy rule extraction, thereby tried the same on numerous frequently used datasets on top of a synthetic dataset through dimensions changing from 4 to 60. They confirmed the efficiency of the process by consuming a tenfold cross-validation setup. The future theme is to focus toward the selection of features with controlled

redundancy. Within the next 10 years utilizing an interval-valued fuzzy rule-based categorization approach, Heinrich et al. [8] developed a classifier to determine the risk of cardiovascular diseases. A previous interval-valued fuzzy rule-based categorization approach and the performance concerning the one is given by two classical fuzzy classifiers were compared. The results reveal that the performance of the latter is statistically better than the former, and in addition, doctors can simply recognize it. Jindal et al. [9] proposed a system to foresee the risk of heart diseases utilizing fuzzy rule-based support system on the basis of Mamdani interface system. The entire performance of the system was assessed by comparing the results with that of neural network and J48 decision tree model. A particle swarm optimization-based approach by Stylianou et al. [10] to train the NN (NN-PSO) classifier was able to predict structure failure of multistoried building. Stylianou et al. [10] proposed a model to forecast the training quality and other relatable mistakes in order to avoid any severe effect on athlete's performance. A novel application of particle swarm optimization (PSO) was implemented by Stylianou et al. [10] in order to separate the patients who are having dengue fever and recovering fast than those who do not have dengue fever. Taylor et al. [11] proposed a novel algorithm by combining neural network along with genetic algorithm to solve the problems of recognition of ISL gestures.

### 3 Research Methodology

Following Fig. 1 represents the methodology adopted. For hospital data, there is an



**Fig. 1** Research methodology

extensive number of missing information which is associated with human errors. Hence, organized information should be filled in the missing areas. Before information imputation, firstly the questionable or deficient restorative information should be recognized and then needs to be adjusted or erased to enhance the quality of information. At that point, data integration is utilized for preprocessing the data.

For the acquisition of useful data, it would be feasible for the doctors to specify a condition and examine whether the symptoms are associated with cardiac diseases. For instance, the doctor can specify the rule as displayed below:

$$\text{Rules: } (\text{Age} > 40) \text{ AND } \left( \text{Blood pressure of } \frac{S}{D} > \frac{120}{80} \right)$$

$$\text{AND } (\text{BMI} > 30) \text{ AND } (\text{Smoking} = \text{high})$$

$$\text{AND } (\text{Alcohol consumption} = \text{high}) \text{ AND } (\text{LDL} > 130 \text{ mg/dL})$$

where, in  $\frac{S}{D}$ ,  $S$  refers to systolic blood pressure and  $D$  refers to diastolic pressure. BMI denotes the body mass index, and LDL represents low-density lipoproteins. The rule mentioned above is a sample set to examine whether the symptoms provided are associated with cardiac diseases, based on the rule mentioned a model was implemented using MapReduce programming.

### 3.1 Algorithm Implementation

**Input: Hospital dataset**

**Output: Whether a patient will get heart attack or not**

Step 1: Inject hospital data into Hadoop distributed file system

Step 2: Apply preprocessing techniques—Fill in missing values

Step 3: I is input on which you want to predict, X and y are training data set, m = No. of training examples, n = No. of features

Mapper()

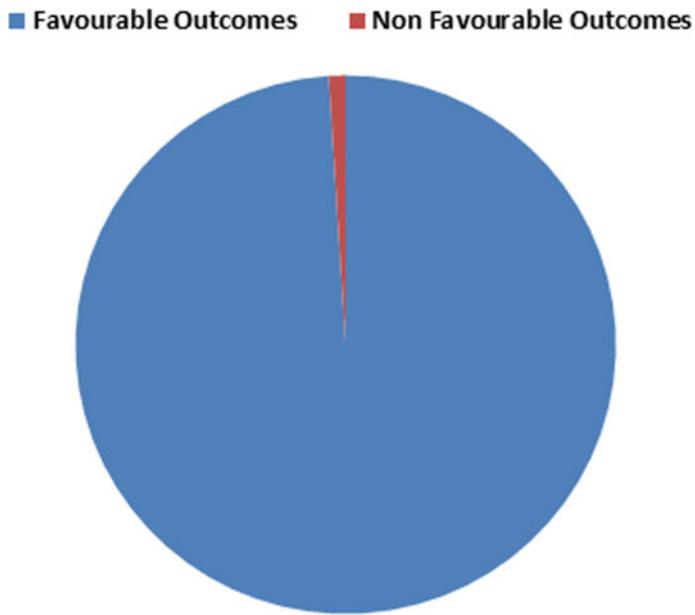
1. function e = execute(I, X, y)
2. [m, n] = size(X);
3. X = [ones(m, 1) X];
4. initial theta = zeros(n + 1, 1);
5. P<sub>a</sub>[N]: Patient's biological parameters
6. j-th biological parameter Health parameter of the patient will be set to abnormal
7. if (labs(P<sub>a</sub>[j] - P<sub>n</sub>[j]) >= c)(c = constant, varying for different health parameters)
8. a\_p = 0, number of abnormal biological parameters
9. Set to true if biological parameter exceeds threshold value
10. **for** j = 1 to N

```
11. if (labs(Pa[j] - Pn[j]) >= c)
12. a_p++;
   Reducer ()
13. I = [ones(1, 1) I];
14. p = predict(theta, I);
15. if p == 1
16. fprintf(['Patient is likely to get a heart attack.\n\n']);
17. else
18. fprintf(['Patient is not likely to get a heart attack.\n\n']);
19. end
```

Above model was implemented using different case scenarios and likelihood is calculated as:

$$\text{Likelihood of an event A} = \frac{\text{Number of Favourable outcomes}}{\text{Total number of possible outcomes}}$$

From Fig. 2, it can be concluded that in 99% of the cases, our algorithm was predicting the correct output.



**Fig. 2** Likelihood of an event

## 4 Conclusion and Future Work

The evolution of internet and computing technologies led to the exponential growth of healthcare applications, wherein the same led to the generation of huge amount of data. Such big data is often stored in cloud environment, wherein the manipulation of such data is deemed to provide useful insights. In this regard, we introduced a mathematical model using MapReduce programming to predict the probability of the heart attack among the patients. However, the researcher suggests future works to be focused on various other nature-inspired algorithms for conditions such as thyroid, diabetes, and so on, wherein datasets with respect to specific developing nation context will be a novel and a useful addition to the research community.

## References

1. Wyber R, Vaillancourt S, Perry W, Mannava P, Folarammi T, Celi LA (2015) Big data in global health: improving health in low- and middle-income countries. *B World Health Organ* 93(3): 203–208. <https://doi.org/10.2471/BLT.14.139022>
2. Dhanalakshmi P, Ramani K, Eswara Reddy B (2017) An improved rank based disease prediction using web navigation patterns on bio-medical databases. *Future Comput Inform J* 2(2):133–147. <https://doi.org/10.1016/j.fcij.2017.10.003>
3. Derhami S, Smith AE (2017) An integer programming approach for fuzzy rule-based classification systems. *Eur J Oper Res* 256(3):924–934 [Online]. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0377221716305240>
4. Gupta N, Ahuja N, Malhotra S, Bala A, Kaur G (2017) Intelligent heart disease prediction in cloud environment through ensembling. *Expert Syst* 34(3):e12207 [Online]. Available from: <http://doi.wiley.com/10.1111/exsy.12207>
5. Suinesiaputra A, Medrano-Gracia P, Cowan BR, Young AA (2015) Big heart data: advancing health informatics through data sharing in cardiovascular imaging. *IEEE J Biomed Health Inf* 19(4):1283–1290 [Online]. Available from: <http://ieeexplore.ieee.org/document/6957068/>
6. Chen M, Hao Y, Hwang K, Wang L, Wang L (2017) Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* 5:8869–8879 [Online]. Available from: <http://ieeexplore.ieee.org/document/7912315/>
7. Chen Y-C, Pal NR, Chung I-F (2012) An integrated mechanism for feature selection and fuzzy rule extraction for classification. *IEEE Trans Fuzzy Syst* 20(4):683–698 [Online]. Available from: <http://ieeexplore.ieee.org/document/6112676>
8. Heinrich A, Lojo A, González AR, Vasiljevs A, Garattini C, Costa-Soria C, Hamelinck D, Artigot EN, Menasalvas E, Xu HF, Sasaki F, Aarestrup FM, Kerremans GR, Thoms J, Sanchez MM (2016) Big data technologies in healthcare: needs, opportunities and challenges. TF7 Healthcare subgroup [Online]. Available from: <http://www.bdva.eu/sites/default/files/Big%20Data%20Technologies%20in%20Healthcare.pdf>. Accessed: 14 Feb 2018
9. Jindal A, Dua A, Kumar N, Vasilakos AV, Rodrigues JJP (2017) An efficient fuzzy rule-based big data analytics scheme for providing healthcare-as-a-service. In: 2017 IEEE international conference on communications (ICC), May 2017, IEEE, pp 1–6 [Online]. Available from: <http://ieeexplore.ieee.org/document/7996965/>
10. Stylianou A, Talias MA (2017) Big data in healthcare: a discussion on the big challenges. *Health Technol* 7(1):97–107 [Online]. Available from: <http://link.springer.com/10.1007/s12553-016-0152-4>

11. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK (2016) Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med Official J Soc Acad Emerg Med* 23(3):269–278 [Online]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26679719>

# Tuning of CNN Architecture by CSA for EMNIST Data



Navdeep Bohra and Vishal Bhatnagar

**Abstract** Convolutional neural network is the deep learning model which has several hidden layers in contrast to feed-forward neural network. Modeling of CNN layers depends upon the dataset and very challenging task as several trials are required to select the CNN parameters. In our work, we presented an optimal solution to tune the hyperparameters of CNN architecture by clonal search algorithm (CSA). This is tested on a challenging dataset of EMNIST, which is enhanced from ML benchmark dataset of NIST. With the proposed algorithm, it is possible to get the accuracy up to 98.7%.

**Keywords** CNN · EMNIST · Clonal search algorithm (CSA)

## 1 Introduction

In the present scenario, data categorization and object recognition is the hot and challenging research area which contributes significantly to robotics and automation. Some previous studies give an idea about training datasets and deep learning networks. The MNIST dataset is considered as the benchmark to test any deep network. It consists of handwritten digits, numeric data, and contains 814,255 characters, which further divides into two categories By-class and By-merge [1]. In a previous study, the clonal search algorithm is proposed for the traveling salesman problem solution. The basic principle of CSA is the human immune system, which

---

N. Bohra (✉)  
USICT, GGSIPU, New Delhi, India  
e-mail: [navdeepbohra@gmail.com](mailto:navdeepbohra@gmail.com)

Maharaja Surajmal Institute of Technology, C-4, Janakpuri,  
New Delhi 110058, India

V. Bhatnagar  
Ambedkar Institute of Advanced Communication Technologies and Research, Geeta Colony,  
New Delhi 110031, India  
e-mail: [vishalbhatnagar@yahoo.com](mailto:vishalbhatnagar@yahoo.com)

contains two essential parameters that are antigen and antibodies. Mutation process occurred when antibodies are selected for the cloning the antigen [2, 3].

Keiron al. discuss the CNN architecture for the image categorization accuracy. The CNN variables are trained using the MNIST database [4, 5]. The CNN is proposed for the image deconvolution purposes. In this method, two models were used for the removal of deconvolution and outliers—the main characteristics of degradation of image achieved by this method [6]. Another efficient classification method is proposed in [7] with a combination of binary CNN and FPGA. The binarized CNN design in the form of the small chip is placed on the FPGA platform. So, FPGA followed the rule of the multiscale sliding window and provided the greater accuracy in object detection. For the classification of handwritten characters, genetic algorithm is used to train the CNN variables. GA developed the optimal structure of the network by tuning the hyperparameters of CNN. This method was tested on the MNIST dataset [8]. A new technique proposed called region-based detection to reduce the complexity and accuracy of object detection. The characteristics of an object recognized by CNN and provided better classification [9]. A convolutional neural network was also proposed for the classification of structural MRI images to diagnose Alzheimer's disease. The ADNI dataset is used for the testing purpose [10]. ANN is used for the recognition of the digit. It is trained by low-level representation data as well as the high-level representation data. In low-level representation, the neural network provided the minimum preprocessing of input data [11]. A general-purpose NN chip can work as an accelerator in large networks.

The literature widely accepts the CNN [12] approach for image and object classifications. We are using the extended MNSIT dataset, which is also image data. We can train the CNN network using the EMNIST dataset and hyperparameters of CNN tuned by the optimization algorithm. In this work, the CSA optimization tunes the CNN network. There are 12 hyperparameters available in the CNN architecture whose values can be optimally tuned for better accuracy. Further, in the paper, EMNIST will be discussed in Sect. 2. CSA optimization is discussed in Sect. 3. Section 4 discusses the methodology of CNN parameters training by CSA optimization. Section 5 describes the results and discussion followed by the conclusion in Sect. 6.

## 2 EMNIST Dataset

The EMNIST dataset is derived from the NIST [1] and MNIST database. The NIST and MNIST dataset has an uneven number of samples with more digit samples than letter samples. So, the MNIST data is extended to provide the balanced samples per class and balanced digit subset to the digit class. The name of the dataset is called extended MNIST or EMNIST. The dataset contains the handwritten samples as collected in the NIST database. The handwritten digits present in the data are used as the testing data information. The EMNIST data has the English alphabets, numeric digits, and the combination of both classes.

The EMNIST database divided into two categories By-class and By-merge. It contains a total of 814,255 characters which distributed in both categories of EMNIST dataset. The digit classes' samples do not change across these two datasets. The digit class of EMNIST dataset contains 28,000 samples of each digit. The EMNIST dataset has a similar size and specification as in MNIST dataset. It can replace the MNIST dataset which has the digit created through the conversion process. It also indicates the validation subset from the training dataset. The samples present in the EMNIST database are a balance in nature and valid for each function [13].

### 3 Clonal Search Algorithm

The natural immune system inspires the clonal search algorithm. The immune system consists of the B-cells and T-cells. Some dangerous foreign cells enter into the human body immune system, which affects them badly known as antigens. Cells induced in the human body which reduces the effects of antigens are called antibody. The recognition among the antigens and antibody considers as affinity. Cellular reproduction occurs due to the presence of B-cells and T-cells, which generate clonal cells, and the process is called clonal expansion. The accelerated mutation process generates a new random population of diverse antibodies. The gene responsible for interaction between antibodies ( $A_b$ ) and antigens ( $A_g$ ) changes randomly, which causes the higher affinity ( $A_f$ ). So, higher  $A_f$  variant selects to enter in the memory cells and plasma, and low  $A_f$  variants are eliminated. If  $A_f < A_g$ , then  $A_b$  removes, and in other case, if  $A_f > A_g$ , then  $A_b$  selects the affinity maturation process. So,  $A_f > A_g$  condition used stimulates to proliferate. Mutation will occur in each proliferates stage. Firstly, initialize both  $A_g$  and  $A_b$  which select the antibodies to identify affinity.

$$A_f = A_g - A_b \quad (1)$$

The number of clones generated can be mathematically written as

$$N_c = \sum_{i=1}^n \text{round}\left(\frac{\beta \cdot N}{i}\right) \quad (2)$$

Here,  $\beta$  is the multiplying factor. Total number of antibodies is  $N$ . Each term of the above formula correlates to the clone size of each selected antibodies.

## 4 Methodology

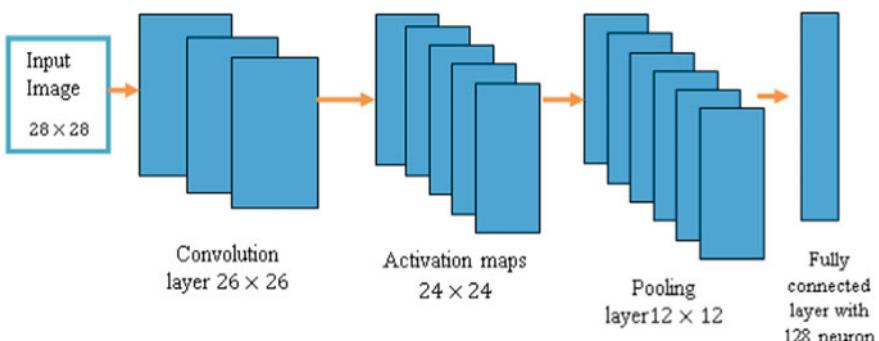
The classification of characters in EMNIST dataset is a machine learning task. The conventional ML algorithm follows the steps of feature extraction, reduction of feature, and training the ML network using activation function. In it, the model only has three kinds of layers mainly and a single type of hidden layers. Deep learning models are evolved from the neural network and suitable for nonlinear and large datasets. The CNN approach is conventional and efficient deep learning model and suitable to train over EMNSIT database.

### 4.1 CNN Architecture

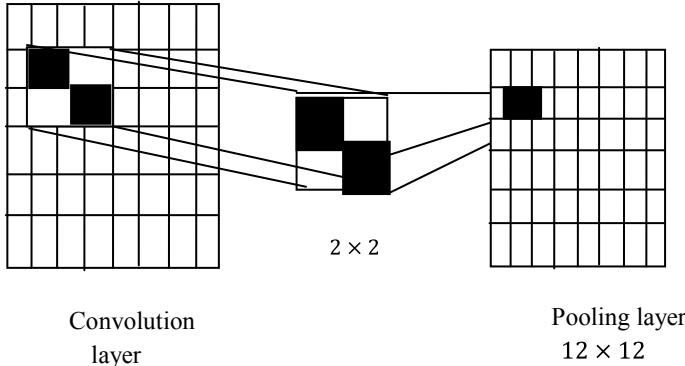
CNN has three main layers known as convolution layer, pooling layer, and fully connected layer. The activation layer is also present in the architecture of CNN, which extracts the features information from the input image. In the CNN model, two operations convolution and sampling are performed. The convolution operation performed in the first layer is known as convolution layer, and sampling function is processed in the max-pooling layer. The final layer performed similar operation like hidden layers in the neural network. So, filter matrix is formed in the convolution layer, and weights are optimized in the fully connected layer. The basic arrangement of CNN layer reflects in Fig. 1. The principle is similar to the regular neural network but different size images provided as the input to the CNN [14, 15].

#### 4.1.1 Convolution Layer Process

The local receptive provides the input to each neuron. The input is extracted from the  $n \times n$  rectangular section which is defined as



**Fig. 1** Architecture of CNN [4]



**Fig. 2** Sampling process in pooling layer [16]

$$x_{i,j}^l = \sigma \left( b + \sum_{r=0}^n \sum_{c=0}^n w_{r,c} x_{i+r, j+c}^{(l-1)} \right) \quad (3)$$

The local receptive fields consider the same weights and biases as formulae in Eq. 3. All the parameters are shown by the trainable filter known as kernel ( $F$ ).

The process of convolution is considered as the acting of image convolution. Figure 2 shows the convolution process with the input layer. The EMNIST database is nonlinear, and convolution function has linear form, so activation function is provided to the CNN model.

#### 4.1.2 Activations Units

The general form of the activation function is described by Eq. 4, where  $x$  is the input image

$$y(x) = \tanh(x) \quad (4)$$

But, this activation unit  $\tanh(x)$  or sigmoid function suffers by the vanishing gradient problem as shown in Eq. 5

$$y(x) = \text{Max } 0, b \sum_{i=1}^k x_i w_i \quad (5)$$

A rectified linear neuron (ReLU) is used to avoid the vanishing gradient problem because it is not saturated for the large input.

#### 4.1.3 Pooling Layer (Sampling Process)

After the convolution process, the activation function is applied and sampling process is achieved between the convolution and pooling layer. It reduces the size of the activation layer output using small sliding window. Pooling layer receives a short rectangular block from the convolution layer and converts into subsamples. Figure 2 shows the sampling process from convolutional layer to pooling layer.

The max function is applied to the activation mapping for scaling the range of activation layer. The extracted features from input layer consist in the activation map. The max-pooling layer receives the best element of ReLU activation map.

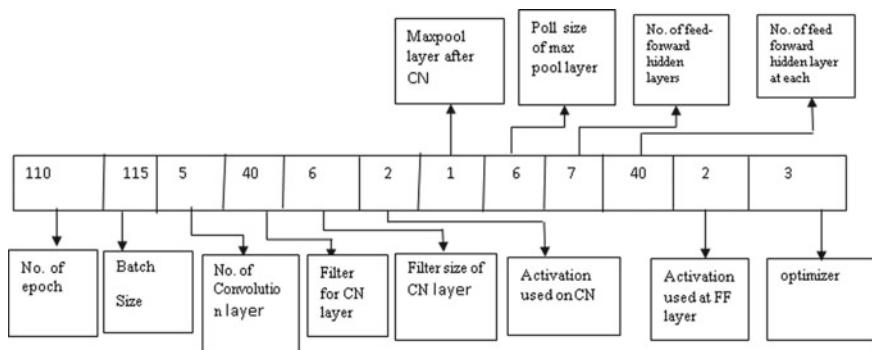
Pooling layer receives a short rectangular block from the convolution layer and converts into subsamples. Figure 2 shows the sampling process from convolutional layer to pooling layer.

The max function is applied to the activation mapping for scaling the range of activation layer. The extracted features from input layer consist in the activation map. The max-pooling layer receives the best element of ReLU activation map.

#### 4.1.4 Fully Connected Layer

The fourth layer shown in Fig. 3 is fully connected layer. It works on the basic principle of a neural network. It contains multiple hidden layers having 128 neurons. Image classification is the key function of a fully connected layer.

The EMNIST database contains the  $28 \times 28$  pixel size gray scale image. In convolution layer, 32 filters are applied to the input image which produces 32 feature map of size  $26 \times 26$ , after convolution layer, the activation layer applied 64 filters to the convolution image which extracts 64 feature maps and produces  $24 \times 24$  size image. The sampling process is followed by max-pooling layer which reduces the image size to  $12 \times 12$  using subsampling window  $2 \times 2$ . The fully connected layer applied the sigmoid function or softmax function with 128 neurons and classified the image.



**Fig. 3** Tuned hyperparameters after the CSA optimization

For the betterment of prediction accuracy, we proposed CSA optimization algorithm for tuning the hyperparameters of CNN.

## 4.2 Tuning of CNN Hyperparameters Using CSA

Structure of CNN layers and its hyperparameters are to be decided before training, and it is done by hit and trial method usually. In our work, we proposed the tuning of hyperparameters for EMNIST dataset, so that an optimal structure with maximum accuracy than conventional CNN layers arrangement can be approached.

The clonal search algorithm tunes the CNN hyperparameters. Size of each layer and its activation function are selected by optimization. A total of 12 hyperparameters are tuned. Table 1 shows the details of those CNN parameters with their specific range.

### Encoding of Hyperparameters into the CSA Clone

The problem should represent in such a way that so it is adaptable for the CSA. The variables involved in the tuning process are the hyperparameters of CNN. The affinity in the CSA is the accuracy of classification which is calculated by using a tuned set of hyperparameters as

$$\text{Accuracy} = \frac{\text{total number of correct predictions}}{\text{total number of predictions}}$$

**Table 1** Hyperparameters range of CNN

Hyperparameter	Range	Hyperparameter	Range
Number of epoch	0–127	Maximum pool layer after each convolution layer	True, False
Batch size	0–256	Pool size of each max-pool layer	0–8
Number of convolution layers	0–8	Number of feed-forward hidden layers	0–8
Number of filters at each convolution layer	0–64	Number of feed-forward hidden neurons at each layer	0–64
The filter size of each convolution layer	0–8	Activation used at each feed-forward layer	Sigmoid, softmax, tanh, ReLU
Activations used at each convolution layer	Sigmoid, ReLU, tanh, linear	Optimizer	Adagrad, Adadelta, RMS, SGD

Antigen and antibodies in the CSA are equivalent to parameters values of CNN layers. An  $A_b$  is represented by a vector of  $1 \times 12$  whose values lie in between the range as in Table 1. We have considered 10 numbers of antibodies in our experiment which makes a total matrix of size  $10 \times 12$  whose initial values are chosen randomly with Gaussian distribution;

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Big|_{\mu=0, \sigma=1}$$

Figure 3 shows a vector of  $A_b$ 's positions specified with these twelve hyperparameters.

Each matrix of antibodies' position sets the hyperparameter and trains the CNN for EMNIST image dataset. Since the EMNIST data has numeric and English alphabets, these are decoded to make a numeric labels matrix. Every ML algorithm uses quantitative labels not qualitative. So, we encoded the English alphabets into corresponding ASCII values as  $A = 65 \dots Z = 90; a = 97 \dots z = 122$ . All hyperparameters in Table 1 are also not quantitative. So, we decoded them into digits like activations function at convolutional layer are four in options, and optimization will choose one of them at a time. We assigned them a new level from 1 to 4, notation is given in Table 2.

The pairing of pool layer with convolutional layer is also encoded to 1 and 0 instead of true and false. Similarly, the optimizer functions at fully connected layers are also decoded as in Table 3.

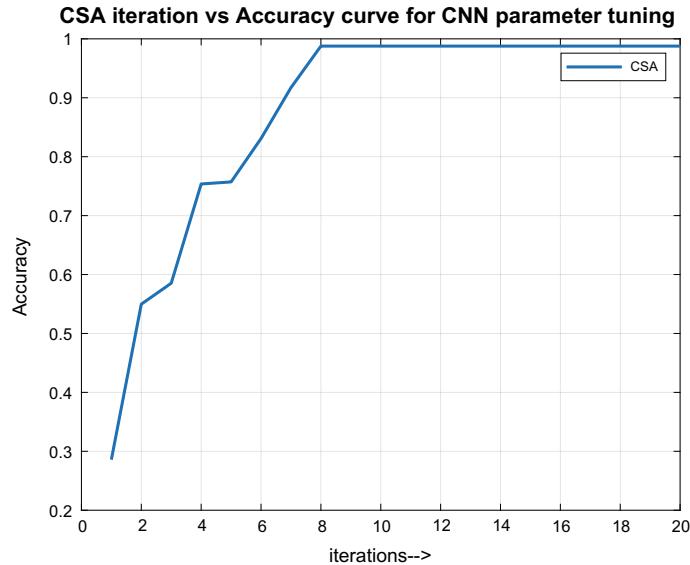
The fitness function for hyperparameters tuning is the maximization of accuracy in each iteration. It is  $\arg \max \text{Accuracy}\{A_b = \{p_k\}_{k=1,2,\dots,K} | K=12\}$ . For each antigen and antibody, it is evaluated and arranged in decreasing order to remove the weakest antigen. New clones are produced by Eq. 2, and for every new clone, fitness function is evaluated again for new accuracy value by trained CNN for these clones. This step is repeated till convergence criteria are reached. For the ideal optimization algorithm,

**Table 2** Encoded activation function

Parameter name	Representation
Sigmoid	1
ReLU/softmax	2
Tanh	3
Linear	4

**Table 3** Encoded optimizer functions for CSA tuning

Optimizer	Symbol
Adagrad	1
Adadelta	2
RMS	3
SGD	4



**Fig. 4** Convergence curve for the hyperparameters tuning of CNN by clonal search algorithm

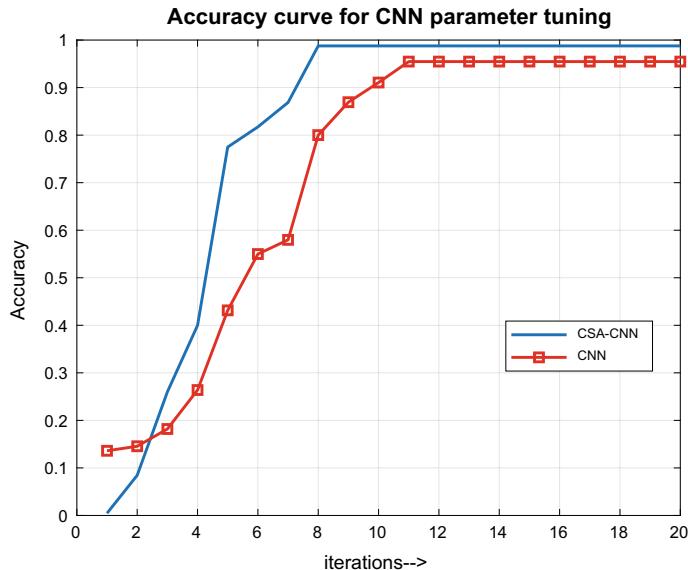
the convergence curve must be increasing with increase in the number of iterations and should settle at an iteration. Earlier it settles, better is the optimization. Figure 4 shows the convergence curve of CSA optimization for this case. It can be observed that accuracy is increasing with every iteration.

The hyperparameters which are tuned to their best values generate the best CNN structure. The best CNN structure gives the highest classification and prediction accuracy for the available EMNIST dataset.

## 5 Results

The CNN parameters are optimally selected for better accuracy in training on challenging EMNIST dataset. We developed the model of the proposed solution in MATLAB and compared the accuracy obtained in recognition of letters' images with conventional CNN with 32 kernels of size  $5 \times 5$  and two cell's padding at the convolution layer. ReLU follows the convolutional layer and max-pooling layer with  $3 \times 3$  pool which down sample the nonlinear features extracted at above two layers. With this configuration, conventional CNN was able to get the accuracy up to 95.4%, whereas after optimizing the parameters by CSA, it reached up to 98.7%. The comparative convergence curve is shown in Fig. 5.

The final-tuned parameters of CNN architecture are given in Table 4.



**Fig. 5** Comparison of EMNIST classification accuracy calculated from CSA tuned CNN model and conventional CNN model

**Table 4** Final-tuned hyperparameters of CNN model by CSA optimization

Hyperparameter	Final values	Hyperparameter	Final values
Number of epoch	78	Maximum pool layer after each convolution layer	True
Batch size	125	Pool size of each max-pool layer	$7 \times 7$
Number of convolution layers	6	Number of feed-forward hidden layers	4
Number of filters at each convolution layer	25	Number of feed-forward hidden neurons at each layer	45 max
The filter size of each convolution layer	$4 \times 4$	Activation used at each feed-forward layer	Softmax
Activations used at each convolution layer	ReLU	Optimizer	Adagrad

## 6 Conclusion

This paper focuses on the hyperparameters tuning of CNN architecture by clonal search algorithm. The CNN configuration changes with each database, and several trials must be done to finalize the number of layers and other parameters. This process even does not assure the highest accuracy. CSA optimization tunes the architecture for any dataset and gives an optimal set of parameter values. The accuracy achieved

by the proposed method is 98.7% which is 3.4% higher than conventional architecture for challenging EMNIST dataset. Our proposed method is also adaptable to every uniform dataset.

## References

1. Grother P (1995) NIST special database 19 handprinted forms and characters database. National Institute of Standards and Technology
2. Muthreja I, Kaur D (2018) A comparative analysis of immune system inspired algorithms for traveling salesman problem. In: International conference on artificial intelligence, pp 164–170
3. de Castro LN, Von Zuben FJ (2002) Learning and optimization using the clonal selection principle. *IEEE Trans Evol Comput* 6(3):239–251
4. Yim J, Ju J, Jung H, Kim J (2015) Image classification using convolutional neural networks with multi-stage feature. In: Advances in intelligent systems and computing, vol 345. Springer, Cham
5. Jaswal D, Sowmya V, Soman KP (2014) Image classification using the convolutional neural network. *Int J Advancements Res Technol* 3(6)
6. Xu L, Ren JS, Liu C, Jia J (2014) Deep convolutional neural network for image deconvolution. In: 27th international conference on a neural information processing system, pp 1790–1798
7. Nakahara H, Yonekawa H, Sato S (2017) An object detector based on multiscale sliding window search using a fully pipelined binarized CNN on an FPGA. In: 2017 international conference on field-programmable technology (ICFPT), Melbourne, VIC, pp 168–175
8. Bhandare A, Kaur D (2018) Designing convolutional neural network architecture using a genetic algorithm. In: International conference of artificial intelligence, pp 150–156
9. Bappy JH, Roy-Chowdhury AK (2016) CNN based region proposals for efficient object detection. In: IEEE international conference on image processing (ICIP), Phoenix, AZ, pp 3658–3662
10. Farooq A, Anwar S, Awais M, Rehman S (2017) A deep CNN based multi-class classification of Alzheimer's disease using MRI. In: IEEE international conference on imaging systems and techniques (IST), Beijing, pp 1–6
11. Le Cun Y et al (1989) Handwritten digit recognition: applications of neural network chips and automatic learning. *IEEE Commun Mag* 27(11):41–46
12. Zhao R, Song W, Zhang W, Xing T, Lin J-H, Srivastava M, Gupta R, Zhang Z (2017) Accelerating binarized convolutional neural networks with software-programmable FPGAs. In: ISFPGA, pp 1–10
13. Cohen G, Afshar S, Tapson J, van Schaik A (2017) EMNIST: an extension of MNIST to handwritten letters, pp 1–10
14. Rastegari M, Ordonez V, Redmon J, Farhadi A (2016) XNOR-Net: ImageNet classification using binary convolutional neural networks, pp 1–55
15. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 39(6):1137–1149
16. Jmour N, Zayen S, Abdelkrim A (2018) Convolutional neural networks for image classification. In: 2018 international conference on advanced systems and electric technologies (IC\_ASET), Hammamet, pp 397–402

# Efficient Emergency Message DHC Broadcasting in Vehicular Ad Hoc Networks



Jaipal, Dhanroop Mal Nagar, and Vinay Baghela

**Abstract** Vehicular ad hoc networks have many access points for communication, transmission, and collecting information of nodes and environment for organization traffic loads. In this paper, we study emergency messaging connectivity in vehicular ad hoc networks (VANETs). Intelligence transportation system applications include two types applications. Comfort applications providing information to the driver about weather, maps and directions, locations, and safety applications are crucial for safety of the driver. This includes such as emergency warning, lane-changing assistance, intersection coordination, which are provided by inter-vehicle communication. Higher No of road accidents demands an intelligent transportation system. Implanting smart sensors, communication capabilities, memory storage, and information processing units in vehicles help us to design an intelligent transportation systems. We are studying the effect of delay of emergency message dissemination. Broadcasting is the familiar way of emergency message dissemination in VANETs. To reduce the broadcasting storm problem and improve scalability of VANET networks, we use a double-head cluster-based broadcasting mechanism. In this research paper, we study the broadcasting delay emergency message of VANET network. The minimum cluster size that achieves acceptable message delivery latency is provided. The simulation results matched those of the analytical model, which showed the analytical model developed in this paper is effective and efficient.

**Keywords** VANET · Broadcasting · Emergency message · Intelligent transportation systems clustering · Double head · Traffic density

---

Jaipal · V. Baghela

Software Engineering, Government Engineering College Bikaner, Bikaner, India  
e-mail: [jaipalbishnoi07@gmail.com](mailto:jaipalbishnoi07@gmail.com)

V. Baghela

e-mail: [baghelavinay@gmail.com](mailto:baghelavinay@gmail.com)

D. M. Nagar (✉)

Information Technology, Government Engineering College Bikaner, Bikaner, India  
e-mail: [drm\\_nagar07@yahoo.com](mailto:drm_nagar07@yahoo.com)

## 1 Introduction

VANET networks are appearing new technology to join the new-generation wireless networks to vehicles. VANET is a special class of mobile ad hoc network, which contains vehicles communicating together between online board units or infrastructure access points—the road side units. It gives wide ranging connectivity while on the road to mobile users, and efficient vehicle-to-vehicle communication between moving vehicles and can be used to design emergency services and traffic management that empower the intelligent transportation systems (ITS).

Vehicular ad hoc networks have access points to access and shared at over the network. Access point can be used to perform communication, transmission, and collecting information of the nodes and environment handling the traffic load in VANET. Vehicles join groups without previous details of other members of the group, after joining and authentication process new Vehicles become a member of the group [1].

ITS applications include two types, the comfort applications and safety applications. Comfort providing information to the driver about weather, maps and directions, near by information locations services, Internet access, and multimedia applications can be provided by vehicle-to-infrastructure communication. On the other hand, safety applications are crucial for ensuring level of safety of the driver. This includes applications such as emergency warning, control of traffic, crash prevention, and real-time path which are mostly to be provided by inter-vehicle communication [2].

## 2 Literature Survey

Algorithms such as SBCA [3], AMACAD [4], and FLBA [5] proposed a dual cluster head (CH) back up that would in some condition stake on the function of the present CH. The prevalent characteristic among these algorithms seems to be that, unless the present CH leaves the cluster, the substitute CH has no position in the cluster. The secondary head used for the DHC, on the other hand, always works to react to any cluster member CM that temporarily loses its link to the main CH.

Two clusters in [6] and [7] must be linked to each cluster throughout the network, which has a comparable concept to use double-head congregating. While this strategy also strengthens accessibility, moreover, the amount of CHs is significantly big and the limits of that same cluster become uncertain. Multi-head clustering had been implemented in [8] in a certain framework including a CH cluster does indeed have a single interface CH and a few slave CHs that are distributed uniformly in the cluster area. The presence of multiple CHs in a cluster was claimed to amplify the stabilization or even boost the cluster's period [9].

Although the idea of establishing a cluster with far more than one CH is partly shared with [8], our algorithm varies in the manner, the secondary CH is chosen and

the main CH is incorporated. The literature used different techniques of CH choice. The common techniques of choice can be categorized into competitive and comparison-based CH selection techniques. UFC and TB [3, 8] are some examples of clustering algorithms that use the rivalry to identify CH [10].

### 3 Pseudocode of Proposed Double-Head Clustering EEMBB Mechanism

Step 1: Check the coordinates of danger vehicle.  
Step 2: Broadcasting the emergency message.  
Step 3: Determining the vehicle state.  
Step 4: Generating a Hello message.  
Step 5: Preparing a Handshake through sending messages.  
Step 6: Clustering and clusters are created.  
Step 7: Process of selecting clusterhead.  
Step 8: In this step, every vehicle joins the cluster.  
Step 9: Calculating the distance and position of last and first CH.  
Step 10: Determine the threshold value.  
Step 11: Distribution of clusters.  
Step 12: In this step, we check the fitness of the cluster and give out fitness factor.  
Step 13: Determine if cluster have sufficient number of vehicles; if yes, end message or remove.  
Step 14: Sending and receiving messages.  
Step 15: Rebroadcasting and maintenance and exit from the cluster.

### 4 Simulator OMNET++ Background

Network simulators are used for real-world communication networks and OMNeT++ is the most reliable and extendable simulation framework by just being a component-based C++ simulation library but not a network simulator by origin. Network simulation can be protracted as, “scheming the synergy between the multiple network entities.” The network itself can be considered as an assortment of several data-sharing devices like main frames, servers, LAN’s, MAN’s, etc. [4]. OMNeT++ expedite its visualization as well as debugging in models of simulation which is easy to customize and embed. It’s modules connect through message passing. OMNeT++ is used in applications due to it’s immense GUI support and modular construction [11].

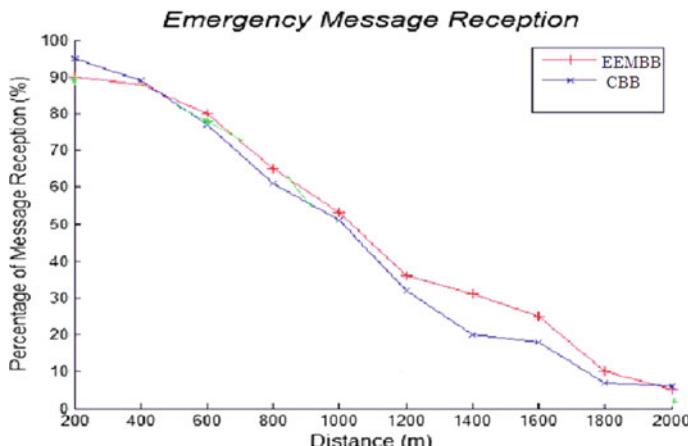
## 5 Simulation Results

This results section presents the broadcasting delay based on the model to enhance emergency message dissemination in VANET; two protocols have been implemented, namely CBB and EEMBB (an improvement, of the Efficient emergency message of CBB). We consider a 512-byte-size emergency message for broadcasting to all the vehicles of the source vehicle [5].

### 5.1 Simulation Results 1

We compared with the CBB and EEMBB. The results of the experiment are shown in Figs. 1, 2, and 3. The generated results focus on emergency message function, channel collision, message reception, and the emergency message delay.

In Fig. 1, CBB and proposed EEMBB protocols results are displayed. The results have been tested in emergency message reception; afterwards, their performances are compare with the CBB protocol. The results show that all the protocols can increase the performances of emergency message acceptance. When SCH (Second CH) connected then broadcasting message becomes more strong. And cover a large areas [12]. Another difference between EEMBB and CBB is that after several tries, EEMBB never fails to rebroadcast the emergency message due to SCH. The EEMBB protocol selects cluster head more carefully because EEMBB depends on traffic and analysis made by neighboring vehicles. We use the algorithm, which takes so the vehicles' analysis into consideration for a perfect selection of preselected cluster head [3].



**Fig. 1** Emergency message receiving with the distance

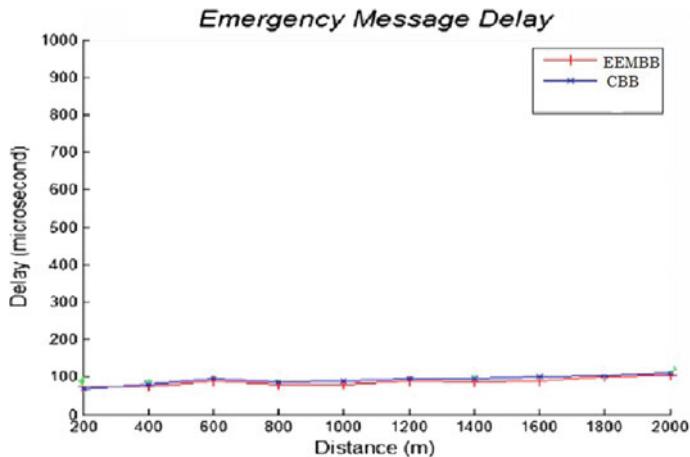


Fig. 2 Emergency message delay message with the distance

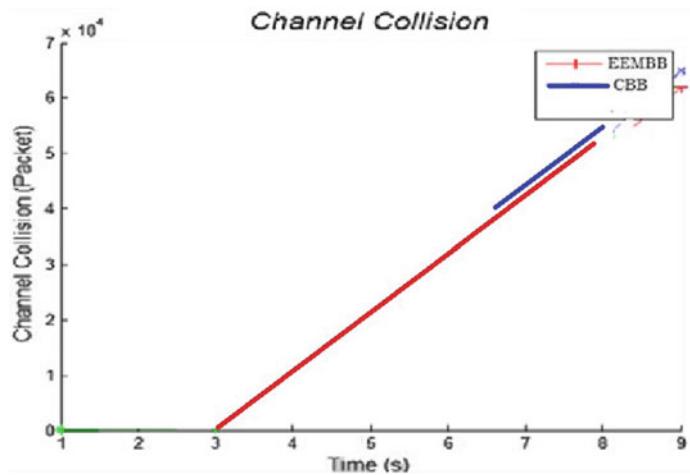


Fig. 3 Collision measured after sending the emergency message

## 5.2 Simulation Results 2

In Fig. 2, we compared the emergency message delay between CBB with EEMBB with distance. In this simulation process, we compute the delay for broadcasting and rebroadcast casting of each unique message, then present that the CBB has a little higher delay than EEMBB during the time but not getting 50  $\mu$ s. If the EMMBB has a less delay at starting point, it means that its rebroadcast effectiveness and decisions are more earlier than CBB. EEMBB has a tiny shorter delay about 10  $\mu$ s shorter than

CBB; EEMBB is an smart technique that has quick response. In safety applications like VANET, a few micro seconds are dangerous in saving life or danger [13].

### 5.3 *Simulation Results 3*

Figure 3 shows the collision generated when broadcasting emergency information from both protocols. At the beginning of the experiment, collisions produced by CBB and EEMBB are noted. However, when we sending a large number of emergency messages resulted in an increase in the number of collisions for both protocols and the difference between the reaching 1% at then in the second.

## 6 Conclusions

ITS focus on reliable and fast transmission of safety messages among vehicles moving in an ad hoc network for some years now. How fast should a message be delivered to each vehicle in the danger zone? Are the communication protocol parameters set to adhere to the delay requirements? This research work indicating a certain perspective to provide answers the above questions.

This paper research work is cussed with two aspects in VANETs. The first is a broadcasting delay of emergency messages in cluster vanet network. This is utilized by highway traffic models developed which analyzing communication delay. We gain expressions for the broadcasting delay based on traffic density regions in the traffic modeling. The second aspect is the study of delay for emergency messaging in VANETs. We investigate the effect of traffic flow density on the total broadcasting delay. Our proposed double-head clustering algorithms build on the clustering infrastructure, introduce the selection process of the clustering member and cluster head. In the domain area, we use double cluster head algorithm to launch the routing path [14].

## References

1. Ishtiaq A et al (2019) Intelligent clustering using moth flame optimizer for vehicular ad hoc networks. *Int J Distrib Sens Netw* 15(1):1550147718824460
2. Haouari N, Moussaoui S, Senouci S-M (2018) Application reliability analysis of density-aware congestion control in VANETs. In: Proceedings of 2018 IEEE international conference on communications (ICC), Kansas City, MO, 20–24 May 2018. IEEE, New York, pp 1–6
3. Nasr MMM, Wang ZG, Shen LF (2016) VANET clustering based routing protocol suitable for deserts. *Sensors* 16(4):1–23
4. Kong X, Li M, Ma K, Tian K, Wang M, Ning Z, Xia F (2018) Big trajectory data: a survey of applications and services. *IEEE Access* 6:58295–58306

5. Jeong S, Baek Y, Son SH (2016) A hybrid V2X system for safety-critical applications in VANET. In: 2016 IEEE 4th international conference on cyber-physical systems, networks, and applications (CPSNA), pp 13–18
6. Lyu F, Zhu H, Zhou H, Zhang N, Li M, Shen X (2018) A novel time slot-sharing MAC for safety messages broadcasting in VANETs. *IEEE Trans Veh Technol* 67(4):3586–3597
7. Abuashour A, Kadoch M (2017) Performance improvement of cluster-based routing protocol in VANET. *IEEE Access* 5:15355–15371
8. Ramakrishnan B, Selvi M, Nishanth RB, Joe MM (2017) An emergency message broadcasting technique using transmission power based clustering algorithm for vehicular ad hoc network. *Wireless Pers Commun* 94:3197–3216
9. Cheng X, Chen C, Zhang W, Yang Y (2017) 5G-enabled cooperative intelligent vehicular (5GenCIV) framework: when benz meets marconi. *IEEE Intell Syst* 32(3):53–59
10. Mylonas Y, Lestas M, Pitsillides A (2015) Speed adaptive probabilistic flooding for vehicular ad hoc networks. *IEEE Trans Veh Tech* 64(5)
11. Liu J, Wan J, Jia D, Li D, Hsu C-H et al (2017) High-efficiency urban traffic management in context-aware computing and 5G communication. *IEEE Commun Mag* 55:34–40
12. Zhang D-G, Liu S, Zhang T, Liang Z (2017) Novel unequal clustering routing protocol considering energy balancing based on network partition & distance for mobile education. *J Netw Comput Appl* 88:1–9
13. Bitam S, Mellouk A, Zeadally S (2015) VANET-cloud: a generic cloud computing model for vehicular ad hoc networks. *IEEE Wirel Commun* 22(1):96–102
14. Cheng JJ, Cheng JL, Zhou MC et al (2015) Routing in internet of vehicles: a review. *IEEE Trans Intell Transp Syst* 16(5):2339–2352
15. Zheng J, Wang Y (2018) Connectivity analysis of vehicles moving on a highway with an entry and exit. *IEEE Trans Veh Technol* 67:4476–4486

# Software Effort Estimation Using Machine Learning Techniques



Ripu Ranjan Sinha and Rajani Kumari Gora

**Abstract** The product/software effort/cost-estimation techniques are applied to predict the effort required to finish the project. An incorrect estimation leads to increase in deadline and budget of the project which may further consequence to failure of the project. The estimation models and techniques are used in different phases of software engineering like budgeting, risk analysis, planning, etc. The effort estimation must be done meticulously in SDLC to avoid any slippage to timelines and over budgeting problems. Techniques of effort estimation can be grouped into two categories, i.e. parametric/algorithmic and non-parametric/non-algorithmic models. To overcome the limitations of algorithmic models, non-algorithmic methodologies have been explored which are based on soft-computing methods. Non-algorithmic techniques include Parkinson, expert judgement, machine learning (ML) and price to win. The ML models have been introduced to handle the flaws of parametric estimation models. These models also complement the modern project development and management. Neural networks, fuzzy logic, genetic algorithms, case-based reasoning, etc., are part of the non-algorithmic models. This review paper focuses on software effort estimation techniques based on machine learning techniques, their application domain, method to calculate software cost estimation and analysis on existing ML techniques to explore possible areas of further research.

**Keywords** Software cost estimation · Non-algorithmic · Algorithmic · Machine learning

---

R. R. Sinha (✉) · R. K. Gora  
Rajasthan Technical University, Kota, India  
e-mail: [drsinhacs@gmail.com](mailto:drsinhacs@gmail.com)

R. K. Gora  
e-mail: [rajanigora@gmail.com](mailto:rajanigora@gmail.com)

R. R. Sinha  
SS Jain Subodh College, Jaipur, India

R. K. Gora  
Computer Science, DCE, GoR, Jaipur, India

## 1 Introduction

In software engineering, the software cost estimation is a key process to produce a quality product. The cost estimation is the process which does not give exact details of effort/cost factors to develop a product/project due to too many variables like human resource, technical environment, etc. **The main reason behind project failure as per**

The International Society of Parametric Analysis (ISPA) [1]	<ul style="list-style-type: none"> <li>“1. Lack of estimation of the staff’s skills and levels</li> <li>2. Lack of understanding the requirements</li> <li>3. Improper software size estimation</li> <li>4. Inability to identify an appropriate software development environment”</li> </ul>
Standish Group International [2]	<ul style="list-style-type: none"> <li>“1. Uncertainty of system and software requirements</li> <li>2. Unskilled estimators</li> <li>3. Budget limitation</li> <li>4. Optimism in software estimation</li> <li>5. Ignoring historical data</li> <li>6. Not realistic estimation”</li> </ul>

The software cost estimation processes are included on each level of SDLC like project planning, project management, scheduling, resource allocation, contract negotiations, etc., to avoid cut-off points of time and budget. The estimation done in initial phases of project should have been done with great precision as any minor deviation in estimation leads to huge deviation in effort estimation for the project which shall have further budgetary and time deadline consequences. Hence, accurate projection of effort/cost has enormous importance which should be done before starting of an SDLC phase. A recurring re-evaluation of efforts should be done to determine the progress [3].

The machine learning techniques consume information of past projects and make a model which is applied to anticipate cost/effort of new project. It is inappropriate to apply one machine learning technique on various project domains, as one function optimally in a domain and other one function better in another domain. As per research carried out in machine learning techniques for cost estimation methods, these techniques have outperformed traditional cost estimation methods. For software cost estimations, many ML methods have been studied like radial basis functions (RBF) neural networks, support vector regression (SVR), case-based reasoning (CBR), MLP neural networks, modified genetic algorithms and multiple additive regression trees.

Although, ML techniques are having issues like availability of data set of current methodology and languages, most of the data sets are of software projects which followed the waterfall model. Currently, Agile methodology is in limelight. The lack of availability in required data set is impacting negatively on training and testing of the techniques which further resulting in degradation of quality of estimated efforts.

Also, fourth- and fifth-generation languages are in dominance. The availability of language specific data sets is also creating roadblocks for the ML techniques.

## 2 Approaches for Cost Estimation

Software cost/effort estimation methodologies are mainly classified into two categories as algorithmic and non-algorithmic based on soft computing.

### **Algorithmic (Parametric) Techniques**

Algorithmic techniques use mathematical formulas based on historical data or theory for cost estimation. These techniques take inputs like source line of code, function points, and other cost drivers, which are tough to get during primal states of SDLC. Examples of algorithmic techniques are COnstructive COst Model (COCOMO), COCOMO II, Function point, Putnam's software life cycle model, Walston-Felix model, Bailey-Basil model etc. These techniques are incompetent to model complex relationship between variables, cannot handle categorical data, and do not have reasoning capability. These techniques are appropriate for specific type of environment. These techniques cannot take decision, draw a conclusion based on the available data, and do not support technological advancement. These shortcomings of algorithmic techniques led to use of non-algorithmic methodologies which are based on soft-computing.

### **Non-algorithmic Techniques Based on Soft-Computing**

Non-algorithmic techniques are capable to take decision and draw a conclusion based on the available data. These techniques have reasoning capability and large knowledge base. In non-algorithmic methods, it has been find out while in algorithmic methods it has been compute. Nowadays, machine learning methodologies are being exercised along with or as a replacement of algorithmic techniques. Expert judgement, price to win, estimation by analogy, Parkinson, neural networks, regression trees, genetic algorithm and neuro-fuzzy are examples of non-algorithmic techniques.

### **Expert Judgement**

It is a human-intensive approach and is based on the practical knowledge of the estimator to provide the estimates for a project based on their exposure on similar kind of projects like Delphi, Wideb and Delphi and Work Breakdown Structure (WBS). It is useful in case of unavailability of quantified and empirical data. It can be used to factor in the difference between past project and proposed project and how these differences are being impacted by new technologies, applications and languages. Although the estimates shall be as good as the experts knowledge and experience, it is a very subjective method and may lack standardization.

### **Top Down**

It may be also called as macromodel which gives overall holistic view of the product and then starts to dive into low-level information like Putnam model. It may be used

for early cost estimation at the time of project inception as there is not any detailed information available. It requires minimal project/product details and focuses on system-level activities; hence, usually faster and easier to implement. Although it provides the basic estimations with not so much details which may deviate to much extent when low-level components shall be analysed.

**Bottom Up** It is opposite of top-down approach. The cost/effort of software modules are estimated and their aggregated results evolve the final cost of the project/product like COCOMO. The estimations are more stable and accurate with all the required details, although it is time consuming to get the detailed cost estimation before the kick-off of the project.

### **Estimation by Analogy**

This technique uses the actual details of past projects and compare with details of proposed project which is of same application domain to derive the cost estimates. It is a data-intensive approach. The estimates are more real as derived from actual projects, although the accuracy of estimates of this model depends on the performance of previous projects. It cannot be applied in case of unavailability of comparable projects.

### **Price to Win**

This estimation technique is based on the estimates in terms of price which are required to win/fulfil the project/contract. It is majorly used for budgeting purposes. Although if not estimated properly, time and money may run out before the completion of project/contract.

### **Parkinson**

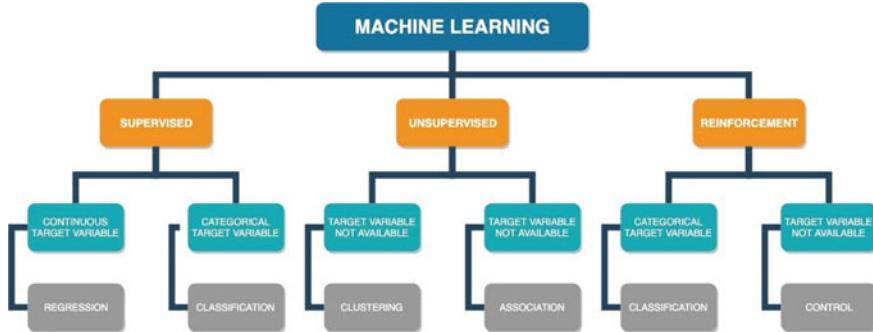
It uses resources as cost estimation rather than objective assessment. In particular scenarios, it provides good results. But usually, it gives unrealistic estimates. Currently, it is not in trend.

### **Machine Learning**

In Machine learning algorithms, a model is created using training data and when new input data is given to the machine learning algorithm, predictions are made based on model. These techniques can deal with real life and ambiguous situations. Machine learning-based techniques are mainly classified into three categories which are shown in Fig. 1.

#### **Supervised Learning** (predictive task)

It is a guided learning procedure. In these learning mechanisms, data sets act as a teacher to train the model. Once the training was completed, then the model can be used to predict or decide on the new data. It can be further grouped into regression (The aim is to anticipate continuous values, e.g. stock values) and classification (The aim is to anticipate discrete values, e.g. {True, False}, {1,0}) problems.



**Fig. 1** Machine learning-based techniques

**List of commonly used algorithm** as neural networks, support vector machines (SVM), linear regression, decision trees and Naive Bayes.

### Unsupervised Learning (descriptive task)

In such models, the training is done without guidance, i.e. information is neither labelled nor classified. The system groups the information using their structural similarities and differences in relationship between data. It finds patterns and relationships between data without labelling it.

**List of commonly used algorithms:** k-means clustering, association rules.

### Reinforcement Learning

It is a kind of dynamic learning which educate algorithm using an arrangement of punishment and payoff, i.e. reward. In reinforcement learning, there is no solution but the reinforcement agent takes proper action to maximize the output in particular context. The agent learns its behaviour through feedback received from environment.

## 3 Machine Learning Algorithms

### Artificial Neural Network

An ANN is a system to process the information. It's modelling and processing is analogous to biological neural network. ANNs are designed to solve a particular problem. There are large numbers of interconnected information processing elements and elements (neuron) are interconnected with the other by weighted links. Each neuron has its own state which is determined through activation function. The activation function is the function of inputs which neuron receives. On ANN, multiple functions can be applied for input like Tanh, sigmoid, linear and Gaussian.

To train an ANN, a technique using existing data set has to be applied. ANNs can be used to approximate nonlinear functions (to map input to a predetermined output), data classifications, pattern recognition, etc. The ANN is of two types:

- feedback networks that employs recursive loops in their paths.
- Feed-forward networks with no loops in their paths.

The most common topology of ANN is feed-forward networks which are normally symbolized as an input layer, hidden layers and output layer. In special case, if hidden layer is not present, then it is called perceptron. Hence, a perceptron maps an input to an output where relationship between its input values and outcome on output is linear. If required, then multiple hidden layers can be introduced to accommodate the properties of nonlinear relationship between input and output. Various types of feed-forward neural networks with hidden layers exist which is comprised of general regression neural network (GRNN), radial basis function neural network (RBFNN) and multi-layer perceptron (MLP).

### **Decision Tree**

Decision tree methodology is a specialized classification and regression. Its modelling technique is easy to understand and has simple classification process. A decision tree is made up of nodes (root, internal and leaf) and arcs. The decision tree algorithm is a data mining induction technique in which data set of information is partitioned recursively using depth-first OR breadth-first to find the data items of a domain. It is having advantage to predict the patterns with missing values and categorical attributes. Regression trees are used for prediction type problems. Regression trees are commonly used to evaluate the cost/effort of the software in software engineering. A regression tree is created by examining the attributes, and considers the most informative attributes. This algorithm is used where outcome of attribute is real number and continuous.

### **Support Vector Machines (SVM)**

SVM is a supervised machine learning algorithm which can be used for both classification and regression problems. It is capable to recognize patterns in data set having noisy and complex data. However, it is mostly used in classification problems for data classes. It performs classification by finding hyperplane between two data class. SVM can simulate complex nonlinear and linear problems and produce highly accurate results, with noisy data as well, resulting from the absence of local minima and the use of kernels. Although its training process can be time hungry for larger data sets.

### **Generalized Linear Models (GLM)**

GLM is a group of general machine learning models for supervised learning problems, i.e. for both regression and classification. It is a generalization of ordinary linear regression that depending on the appropriate link function, variance, distribution and response variable type, is used to depict the relationship among attributes.

### **Naive Bayes (NB)**

NB is a plain and compelling algorithm for predictive modelling. It is based on Bayes theorem. A NB classifier is a probabilistic machine learning model which is utilized for classification.

### **Case-Based Reasoning**

CBR technique solves a problem using reference to previously successful solutions to the same kind of problems. It collects and stores the results and information of past projects. It is simple and highly effective hence widely studied and implemented. CBR working cycle can be as below:

S. No.	Cycle stage	Description
1	Case retrieval	Once the problem has been analysed then the best-fit case is explored in case base. After finding the best matched case, a solution is extracted
2	Case adaptation	The extracted solution is modified to better fit the new problem case
3	Solution evaluation	The modified solution will be checked either before the solution is applied or after the solution has been applied to the problem. If the outcome is not adequate, then the extracted solution must be modified again or more similar cases should be brought
4	Case-base updating	If the outcome is confirmed as correct, then same case shall be added to the case base

### **Rule-Based Systems**

Rule-based systems are handy in estimating the effort of software. In such a system, a set of relation rules collectively represent the knowledge. It has a knowledge base where rules are stored which has been used to compare the characteristics of proposed projects. It uses IF/THEN statement to implement the system. The control scheme is goal driven if it does backward chaining and data driven if it does forward chaining. The chaining effect continues to be fired until there are no rules to check.

### **Genetic Algorithm**

Genetic algorithm is adaptive heuristic search algorithm and belongs to the evolutionary methods of cost/effort evaluation. Evolutionary computation methodologies are defined by the information that the solution is attained by means of a sequence of generations of candidate solutions that are trimmed by the criteria “survival of the fittest”. It is employed for solving both constrained and unconstrained problems. In GA, for software cost/effort estimation, the algorithm is repeatedly applied on population of solutions to modify. On every iteration, the algorithm selects individuals from population to make them parents and use them to create children of the next generation. Over multiple successive generations, the population evolve to an optimal solution. It uses three major types of rules, i.e. Selection rules to pick out the parents, Crossover rules to make children by combining two parents and Mutation rules for stochastic alterations to parents to make children. GA is used to render high quality solution to search problems and optimization.

S. No.	Machine learning technique	Strength	Weakness
1	Artificial neural network	<ul style="list-style-type: none"> <li>1. It stores information across the network rather than in a database. Loss of few information at intermediate node does not prevent network from functioning</li> <li>2. It can work with incomplete information. However, performance of the network may be degraded depending on criticality of the information</li> <li>3. It has fault tolerance capability as one or more nodes do not prevent it from generating output</li> <li>4. It has parallel processing capabilities</li> </ul>	<ul style="list-style-type: none"> <li>1. As ANN requires processors with parallel computing according to their network structure, the realization of ANN is costlier</li> <li>2. When an ANN gives predicted/unpredicted output, it does not provides HOW and WHY</li> <li>3. There is no standard way to determine the architecture of network. It is to be achieved through experience and trial and error</li> <li>4. An ANN understands numerical information only. Transformation of problem into numerical information depends on ability of user</li> <li>5. There is always an error in result. The size of the error depends on the training data</li> </ul>
2	Decision trees	<ul style="list-style-type: none"> <li>1. Decision trees are easily understandable and easy to use</li> <li>2. Do classification without requiring much computing power as it does not require normalization of data</li> <li>3. Can handle continuous and discrete variables. Missing information of data does not stop the making of the tree. Although tree cannot be made for that consequence</li> <li>4. Can give comprehensive analysis of all possible outcomes</li> </ul>	<ul style="list-style-type: none"> <li>1. Decision trees are not preferred for the prediction problems</li> <li>2. They are prone to errors when count of classes is high but training data sets are less. A small change in data set lead to large change in structure of the decision tree causing instability</li> <li>3. They are expensive in terms of computation power requirement as all possible outcomes has to be draw to see final optimum result</li> <li>5. They are complex and time consuming for large problems. They require higher time to train the model</li> </ul>

(continued)

(continued)

S. No.	Machine learning technique	Strength	Weakness
3	Support vector machines	<ul style="list-style-type: none"> <li>1. It works well with unstructured OR semi-structured data. It has L2 Regularization feature as the risk of over-fitting is less</li> <li>2. Unlike neural net, it doesn't solve for local optima</li> <li>2. It can handle nonlinear data efficiently. It is highly scalable for high dimensional data</li> <li>3. It can solve both classification and regression problem efficiently</li> <li>4. A small change in data does not impact the hyperplane. Hence, it is more stable</li> </ul>	<ul style="list-style-type: none"> <li>1. Selection of a good "kernel" function is a tough task. Also algorithmic complexity and memory requirements are too high</li> <li>2. Long training time is required for large data sets</li> <li>3. Difficult to understand and interpret the final model; it is tough do the small calibrations to visualize the impact. Hence, it is hard to implement the business logic</li> <li>4. Not suitable for data sets which are noisy like overlapping classes</li> </ul>
4	Generalized linear models	<ul style="list-style-type: none"> <li>1. GLMs extrapolate over predictor levels with little or no data</li> <li>2. GLMs provide easily calculated relativities to use as a rate classification system</li> <li>3. By assuming you know the form of the "noise", you can do statistical inference to evaluate predictors</li> <li>4. You can also provide confidence intervals to communicate the inherent uncertainty in the output</li> <li>5. Easy to interpret and understand HOW</li> </ul>	<ul style="list-style-type: none"> <li>1. GLM model risk can be mitigated but not removed</li> <li>2. GLMs simply do not provide a system for finding all of the relevant interactions</li> <li>3. GLMs are not formulated to find local interactions. It uses global interactions</li> <li>4. GLMs involve an extensive modelling process which requires immense time and effort to moderate the modelling process</li> <li>5. GLMs will not give the predicted output if assumptions laid under the model are failed</li> </ul>
5	Naïve Bayes	<ul style="list-style-type: none"> <li>1. They work well on independent variables as compared to other models</li> <li>2. For estimation, small amount of training data needed</li> <li>3. It is easy to understand and implement</li> </ul>	<ul style="list-style-type: none"> <li>1. The limitation is that it works only on independent variable. In real world, it is almost impossible to get variables which are independent to each other</li> <li>2. If any category of training data set is not provided, it will mark as 0 probability for the scenario</li> </ul>

(continued)

(continued)

S. No.	Machine learning technique	Strength	Weakness
6	Case-based reasoning	<ul style="list-style-type: none"> <li>1. It has simple and comprehensive knowledge representation and can express specialized knowledge</li> <li>2. It has modular characteristics as each case is a discrete knowledge unit</li> <li>3. It can handle missing and inconsistent data with the help of self-updatability</li> </ul>	<ul style="list-style-type: none"> <li>1. CBR is usually used for specialized knowledge. Hence, general knowledge can't be expressed</li> <li>2. For unavailable or limited case domains, the inference process hinders due to lack of cases</li> <li>3. It is suffered from inference efficiency problems</li> </ul>
7	Rule-based systems	<ul style="list-style-type: none"> <li>1. General Knowledge about a problem domain can be easily represented. These can be representation of expert systems</li> <li>2. It has a modular characteristics as each rule is a discrete knowledge unit</li> <li>3. The explanation of derived conclusion is easy to interpret</li> </ul>	<ul style="list-style-type: none"> <li>1. The standard way to identify rules from various data collection techniques is not up to the mark to cater the transformation of information to rule</li> <li>2. The final outcome cannot be determined in the case of missing or inconsistent data</li> <li>3. These systems suffer from inference efficiency problems and interpretation problems</li> <li>4. It is hard to maintain the large rule base</li> <li>5. Rule-based systems are not intelligent and are not self-updatable</li> </ul>
8	Genetic algorithm	<ul style="list-style-type: none"> <li>1. It can find local optimal solution in very less time</li> <li>2. The mutation guarantees that most of the solutions can be explored</li> <li>3. It is easy to understand, operate and parallelized</li> <li>4. It works matter in noisy environment</li> </ul>	<ul style="list-style-type: none"> <li>1. It is expensive in terms of computation</li> <li>2. Although Genetic algorithm requires less information, designing and determination of objective function may be difficult</li> </ul>

## 4 Literature Review

Singh et al. [4] used Improved Environmental Adaptation Method with Real Parameter (IEAM-RP) to predict the cost of software development. It is an evolutionary algorithm which is achieved using operators, adaptation and selection and is used to solve single objective optimization problems. The authors stated that these algorithms were successful for predicting efforts resulting from their population-based search techniques. This algorithm minimizes the difference between measure efforts and estimated efforts. The effectiveness of the described method was checked over the NASA software project data set and observed that the estimation of cost/effort needed to make a software system is quite good in terms of population diversity and convergence rate.

Usman et al. [5] did a multi-case study of expert judgment-based techniques for improving software effort estimation. In expert-based estimation techniques, experts may ignore important factors that cause underestimation and many more. To overcome the aforementioned problems, the authors proposed a method to develop an estimation checklist for agile teams and accessed the worth of checklist for expert-based estimation techniques. The authors implemented the proposed method in three different organizations to make and validate checklists in the particular environment. Checklist facilitated several advantages during the estimation process like reliability of estimates, better consistency by recollecting relevant factors, improved understanding of the tasks being estimated, more objectivity in the process and reduce the chances of missing tasks. Authors described that the checklist can be useful in reusing estimation data and documenting which result in better estimation accuracy. The authors concluded that effort estimation was improved when checklists were included in the estimation process.

Ezghari et al. [6] discussed Fuzzy Analogy Based Software Effort Estimation Model (FASEE) for the capability of the model to handle imprecision and uncertainty of the model's reasoning process. The authors observed that FASEE still impacted by low data quality which further causes uncertainty in effort estimation. The author introduced the Consistent Fuzzy Analogy Based Software Effort Estimation Model (C-FASEE) to elevate FASEE's aforementioned weakness. The proposed model introduced two more information. The first one is consistency criteria attribute in the fuzzy set for increased precision and the second one is confidence relation between similar projects for uncertainty quantification. The proposed model is validated using thirteen software data sets and the achieved results were compared with various versions of the analogy-based software effort/cost estimation approach. Authors found that the proposed model has high estimation accuracy against the predecessor.

Qi et al. [7] did a study on available data sets for software effort estimation and found that most of the organizations did not share the project's effort data because of privacy concern which lead a limited amount of effort data. Software effort estimation on limited data gives an unrealistic estimation. Due to this, the authors proposed a method to reduce the shortage of data by choosing GitHub data and also proposed

AdaBoost and Classification and Regression Tree (ABCART), a sample incremental algorithm that increases the samples of collected data sets online and also satisfy the requirement of the ever-changing growth of data sets. Authors concluded that effort estimation for new projects that do not have training data can be easily done by using open source project (OSP) data. Experimental result disclosed that the estimation performed on collected data from OSP has comparable performance with those of existing effort data sets.

Sehra et al. [8] did a study on Research Trends and Patterns in Software Cost/Effort Estimation to identify unobserved research trends prevailing in software effort estimation literature. The authors have carried out a assessment upon 1178 articles on software effort estimation published during the period 1996 to 2016 through natural language processing using Latent Dirichlet Allocation which is statistical method. The authors identified sixty research trends and twelve core research areas which help in finding the future research areas.

Mensah et al. [9] introduced a self-guided interpretation and classification method to improve conclusion instability of Software Effort Estimation (SEE) because conclusion instability has an impact on the enactment of SEE models. The difference in effort estimation results from various SEE models leads to conclusion instability. To validate the output, the authors used leave-one-out cross-validation and analysed using adjusted R<sup>2</sup>, Balanced Mean Magnitude of Relative Error (BMMRE) and mean absolute error (MAE). This study used 14 historical data sets for their experimental purpose. The authors concluded that the above-mentioned approach improve prediction accuracy and minimizes conclusion instability. For the future research point of view, the authors suggested that one can examine the above-mentioned approach upon various repositories or organization with the intention of improving the generalizability of results.

Benala et al. [10] introduced a feature weight optimization technique called Differential Evolution in Analogy-Based Software Development Effort Estimation (DABE) For checking the effectiveness of DABE, the authors have done exhaustive study of PROMISE repository test suite and found remarkable improvements in predictive performance of DABE methodology over ABE. The authors used standardized accuracy (SA) and EF unique global error indicators for measuring performance of their study.

Srdjana et al. [11] proposed Bayesian network model for cost/effort estimation in agile software development which was small and simple. The authors proposed this model because Bayesian network is employed in software development projects to estimate the efforts, reliability evaluation, risk assessment and quality prediction. The authors depicted that the proposed mechanism can be applied in early stages of SDLC and it does not affect the agility. The authors also developed a method for elicitation, documentation and validation of software user requirements (MEDoV). When the above method was applied in an agile project, then the project was finished on time and within the predicated budget. The authors concluded that proposed model has satisfactory prediction accuracy.

Pospieszny et al. [12] proposed an effort and duration estimation model using smart data preparation, a set of three machine algorithm (named as support vector

machine, multi-layer perceptron and generalized linear models) and validated using three-fold cross-validation, whose purpose is to work as a decision support tool for any organization. The authors depicted that this model is suitable for medium and large scale organization which have significant volume of finished projects. The proposed model is more appropriate in the initial stages of software development life cycle where uncertainty of deliverable product is high. The authors concluded that the results imply very good prediction accuracy.

Arau et al. [13] developed a hybrid multi-layer perceptron, known as multi-layer dilation-erosion-linear perceptron (MDELP) for software development effort/cost estimation. Mix of hybrid linear operator and morphological operator was used to compose the processing units of perceptron. To train the proposed model, the authors also proposed a descending gradient-based learning process. The authors concluded that the model had better performance.

Jørgensen [14] did a study on the impact of effort on expert judgement-based cost estimation of software development. For the above-mentioned purpose, two experiments were conducted in which software engineers were stochastically directed to perform cost/effort estimation of a project in work-hours or work-days. The practical consequence of unit effect showed that estimation in work-hours had degraded estimates than work-days in both experiments. The author concluded that if there is a chance of underestimation, then estimation in work-days instead of work-hours gives more realistic estimation.

Idri et al. [15] did a systematic review of Ensemble Effort Estimation (EEE) techniques on 24 selected papers published between 2000 and 2016 in four literature resources (Google Scholar, Science Direct, IEEE Xplore and ACM Digital Library). EEE techniques were introduced to eliminate the weakness of single effort estimation model of effort estimation. Several classical effort estimation models are combined in Ensemble Effort Estimation technique. According to [16], EEE techniques were broadly classified into two categories: heterogeneous and homogeneous. Former EEE combines at least two different base models. Homogeneous combines one base model with at least two different configurations. The authors concluded that the homogeneous methods are commonly found in the literature and the estimation accuracy of EEE mechanisms exceeded the single model.

Zare et al. [17] proposed an updatable three-level Bayesian Belief Network (BBN) based on COCOMO components for software cost/effort estimation. Unlike the BBN in which intervals between nodes are represented by discrete numbers, authors have replaced these discrete numbers with fuzzy numbers to improve the accuracy. The effort estimates were updated by calculating the optimal coefficient which is derived by optical control concept under genetic algorithm (GA). The proposed model also estimates efforts with respect to the number of defects detected and removed in software project life cycle. Estimated effort for the project shall be increased if number of defects has crossed a predefined threshold. The fidelity of the prescribed model is more accurate. It was further suggested to include feedback loops in the estimation process so that the model shall be capable of learning from all historic information.

Idri et al. [18] explored missing data methodology (KNN imputation, deletion and toleration) using three missingness methods, i.e. non-ignorable missing (NIM), missing at random (MAR) and missing completely at random (MCAR) on two methods, namely as classical analogy-based software effort/cost estimation and fuzzy analogy-based software effort/cost estimation. Authors examined their working on seven data sets in terms of standardized accuracy (SA). Authors concluded that when effort estimation is done on missing data, then the output of both analogy-based cost/effort estimation methods had improved using k-nearest neighbour (KNN) imputation method, rather than toleration and deletion mechanisms. For all data sets, fuzzy analogy-based effort estimation outperforms on traditional analogy irrespective of missingness method and the missing data handling methods. In future opportunities, authors suggested to repeat this study on different data sets and experiment was done only on numerical data and further experiments can be done mixed data (categorical and numerical) to ascertain their results, although all three missing data techniques render inferior result for NIM as compared to MAR and MCAR.

Sree and Ramesh [19] had used fuzzy logic controller (FLC) to do effort estimation for software/product but the rule base was large for computation. So, the purpose of this paper was to minimize rule base and enhance the working by cascading of FLC. For this purpose, FLC was cascaded into two and six stages, and by this, the rule base was decreased and the performance was improved as cascading of FLC increased. But the problem with this approach was to identify correct number of cascaded FLC. To overcome this problem, the authors proposed a fuzzy model with properties of subtractive clustering. It reduced rule base and provided better software effort estimation.

## 5 Conclusion

In this review, a comprehensive analysis of software cost estimation techniques based on machine learning has been assessed. The basic goal to conduct the study is to understand the implication of machine learning techniques in the field of software/project cost and effort estimation. In modern project management and development methodologies, rapid change in time-to-market, enormous user requirements and unpredicted changes are quite common. To cope up with these, parametric techniques may be incapable and impractical. The ML techniques for effort estimation are proven to handle deficiencies of existing parametric techniques and give an edge over such issues, although no model is perfect for each software project. Before choosing an ML model with the precise outcome, it is necessary to know about specific characteristics like the availability of required data set of a particular methodology for learning, the kind of input required for ML method, type of programming language, etc. Despite many software cost and effort estimation models, the need for new and improved techniques are required to cope up with the dynamics of information technology. In future work, a comparison between models of different techniques with

different types of data sets may be carried out. Based on the comparison and performance of current techniques, new and improved methods and techniques may be introduced.

## References

1. International Society of Parametric Analysis (2003) *Parametric estimation handbook*, Chap 6. ISPA
2. The Chaos report (1995). The Stadish Group International, Inc. [Online]. Cited: 20 Nov 2009
3. Sinha RR, Gora RK (2018) Review of analysis on selection of different cost estimation models. In: International conference on role of ICT in higher education and research (ICRA), pp 145–147
4. Singh T, Singh R, Mishra KK (2018) Software cost estimation using environmental adaptation method. In: 8th international conference on advances in computing and communication (ICACC-2018)
5. Usman M, Petersen K, Börstler J, Neto PS (2018) Developing and using checklists to improve software effort estimation: a multi-case study. *J Syst Softw* 146:286–309
6. Ezghari S, Zahi A (2018) Uncertainty management in software effort estimation using a consistent fuzzy analogy-based method. *Appl Soft Comput J* 67:540–557
7. Qi F, Jing XY, Zhu X, Xie X, Xu B, Ying S (2017) Software effort estimation based on open source projects: case study of GitHub. *Inf Softw Technol* 92:147–157
8. Sehra SS, Brar YS, Kaur N (2017) Research patterns and trends in software effort estimation. *Inf Softw Technol* 91:1–21
9. Mensah S, Keung J, Bosu MF, Bennin KE (2017) Duplex output software effort estimation model with self-guided interpretation. *Inf Softw Technol* 94:1–13
10. Benala TR, Mall R (2017) DABE: differential evolution in analogy-based software development effort estimation. *Swarm Evol Comput* 38:158–172
11. Srdjana D, Stipe C, Mili T (2017) Bayesian network model for task effort estimation in agile software development. *J Syst Softw* 127:109–119
12. Pospieszny P, Chrobot BC, Kobyliński A (2017) An effective approach for software project effort and duration estimation with machine learning algorithms. *J Syst Softw* 137:184–196
13. Araujo RDA, Oliveira ALI, Meira S (2017) A class of hybrid multilayer perceptrons for software development effort estimation problems. *Expert Syst Appl* 90:1–12
14. Jørgensen M (2016) Unit effects in software project effort estimation: work-hours gives lower effort estimates than workdays. *J Syst Softw* 117:274–281
15. Idri A, Hosni M, Abran A (2016) Systematic literature review of ensemble effort estimation. *J Syst Softw* 118:151–175
16. Elish MO, Helmy T, Hussain MI (2013) Empirical study of homogeneous and heterogeneous ensemble models for software development effort estimation. *Math Probl Eng*
17. Zare F, Zare HK, Fallahnezhad MS (2016) Software effort estimation based on the optimal Bayesian belief network. *Appl Soft Comput J* 49:968–980
18. Idri A, Abnane I, Abran A (2016) Missing data techniques in analogy based software development effort estimation. *J Syst Softw* 117:595–611
19. Sree PR, Ramesh SNSVSC (2016) Improving efficiency of fuzzy models for effort estimation by cascading & clustering techniques. In: International conference on computational modeling and security (CMS 2016). *Proc Comput Sci* 85:278–285

# Sentiment Analysis of English-Punjabi Code-Mixed Social Media Content to Predict Elections



Mukhtiar Singh, Vishal Goyal, and Sahil Raj

**Abstract** On social media, the number of users are increasing exponentially. The information contents posted and tweeted by the user are also increasing exponentially. A different meaning of the sentiment is hidden inside the message. Analysing the nature of the text is still a very challenging task. The sentiment analysis is one of the emerging and challenging fields. In the proposed work, the data has been extracted from Twitter with the dataset of around 145,464 comments. In particular, the English-Punjabi dictionary has been created for opinionated word. The opinionated words are categorized into two parts as positive dictionary and negative dictionary. These are stored in gazetteer list and then a statistical technique has been applied for sentiment analysis.

**Keywords** Sentiment analysis · Tweepy · Text · Twitter · NLTK · Social media content

## 1 Introduction

In the last few years, social media has been a healthy platform for politicians to get views of public at large, and making the data available to their potential voters [1]. During elections, each eligible citizen (above 18) is allotted a vote. There are around more than 30 lakh users surfing on Internet who fall between the age group of 18–35 years and they communicate on social media. The Facebook Web page of Aam Aadmi Party (AAP Punjab) has been liked by nearly ten lakh followers. Many interesting

---

M. Singh (✉) · V. Goyal

Department of Computer Science, Punjabi University, Patiala, India  
e-mail: [mukhtiarrai73@gmail.com](mailto:mukhtiarrai73@gmail.com)

V. Goyal  
e-mail: [vishal.pup@gmail.com](mailto:vishal.pup@gmail.com)

S. Raj  
School of Management and Studies, Punjabi University, Patiala, India  
e-mail: [dr.sahilraj47@gmail.com](mailto:dr.sahilraj47@gmail.com)

videos which are original or in morphed form become a rich source of discussion. One such example is a video in which Kejriwal is assigning third position to Sukhbir Singh Badal. Also rivals like Sukhbir Badal are found liking the same page. The official Web page of Congress, i.e. Captain Amarinder Singh is also has around nine lakh followers. There are number of hashtags that are associated with polling parties; some of them are #AAP-Hi-SADhai (AAP is Badal party), #AAPPunjabVichSaaf (AAP is expected be totally wipedout of Punjab), #SADisBAD (Badal party is terrific). There are Web pages on which unofficial videos showing Captain Amarinder Singh beating Kejriwal and Prakash Singh Badal and others are being posted. The comments on these posts become challenging for opinion mining. In this paper, tweets, retweets of Aam Aadmi Party (AAP), the Shiromani Akali Dal (SAD) and Congress have been extracted using twitter API.

## ***1.1 Sentiment Analysis***

Nowadays, sentiment analysis has been coined as a new term opining mining. It is a very popular area in the field of research. On microblogging websites, considerable amount of textual data is readily available. The textual information is categorized into mainly two parts: Facts and opinions. Facts are subjective expressions about entities, which describe people's feelings towards the entity. Opinions are helpful for decision making related to number of domains like political, agriculture, sports, entertainment, etc. In this paper, the textual data related to the political domain has been extracted from Twitter.

## ***1.2 Sentiment Analysis on English-Punjabi Mixed Language***

In India, there are 22 official languages. On the social media platform, there are 100 millions of Punjabi language speakers worldwide. The geographical area of Punjabi language is also growing day by day via Internet. The Web pages contain a huge amount of important data regarding corporate and government sector which is easily available on various government websites. The need of the hour is to develop resources, approaches and tools for successful research in the Punjabi language, as they are very limited in number [2]. Punjabi is Indo Aryan as well as cultural language. Since few years, there has been a trend of multilingual speakers often switching between more than one language has been noticed on social media network. There has been a strong variation regarding their expressions and opinions. The related concept is known as code mixing. Code mixing means switching from one language to another within the same utterance or written text. Sentiment analysis is a process of analysing emotions, feelings, opinion, and the attitude of a person that is expressed in the form a given input of text information/data. The main aim of sentiment analysis is to identify sentiments associated with the text by extracting sentimental context from

the text [3]. To find the attitude, state of mind, and emotions of individuals through the contextual analysis, sentiment analysis is used. The attitude can be reflected by their judgment, the emotional statement of the opinion, or the statement of any emotional conversation used to influence a reader or listener. To determine an individual's state of mind about the opinion communicated, sentiment analysis is used. The huge amount of data can be fetched from various data resources like texts, twitter data, Facebook data, blogs, social media, news articles, product reviews, etc. [3].

The remaining of the paper is organized as follows: Sect. 2 gives an overview of the background and related work. Section 3 provides proposed work of sentiment analysis. Section 4 depicts the architecture of the system; in which, tokenization, stop words are removed and data pre-processing for classification is described. Section 5 deals with finding the polarity of the data. Last section contains the implementation and results related to the political domain. The conclusion and future work are drawn in Sect. 7.

## 2 Background

Industries are increasing huge amount of investments in marketing sector day by day for collecting the reviews of public towards a specific product [4]. Industries make access to that feedback to improve the quality and manufacturing of their products and customer services as well. The number posts by user's leaves a strong impact on the product revision [5]. Semantic orientation of adjectives and verbs is calculated in a sentence and determined the overall polarity by adding up the independent values for semantic orientation. The dictionaries for positive and negative words and then by integrating negation words, the polarity has been determined. On an average, 68% accuracy has been achieved using this technique. Various strategies are used for inferring political interpretations like profile information, user behaviour, user graph, Twitter-specific feature and sentiment from tweet content [6]. The tweets containing references to several political parties and political events represent the ideological leaning of the user. Following similar methods, the tweets and retweets of a user are used for relating user to a political party and to infer their political leaning [7]. Assigning a score to every member which is being followed by a Twitter user and then assigned a political preference based on that score [8]. Several features such as users posting behaviour, linguistic content, followers, replies and retweets are compared in [9]. It has been found that the combination between the user profile and linguistic outperforms other features [10]. The clustering approach has been used for managing Web news data. This research has used back propagation neural network and k-means clustering for classifying the news data. An attempt to fetch twitter data has been done in [11]. This research used Hadoop distributed file system for storage of data which was fetched with the help of flume. Decision tree and Naïve Bayes classifiers were used for sentiment analysis. The sentiment analysis is done with semantics by training the data from twitter [12]. The authors have used multiple datasets in this work. This research has achieved better results for twitter sentiment

classification in terms of precision a, recall and f-measure. Twitter sentiment analysis is done on dataset of feedback of people related to tourism in Oman [13]. This research also provided a machine learning approach for recommending innovation in sentiment analysis method used. Sentiment analysis is done for creating dynamic profile in voting applications in comparison to static profiles [14]. This research used user-based analysis system. Sentiment analysis is performed with the help of convolutional neural networks in [15]. This research proposed an architecture for performing sentiment analysis of small, medium and large datasets.

### 3 Proposed Work

The main objective of the proposed work is to complete the task of sentiment analysis of predicting the political elections in Punjab. The sentiment analysis for the Political Parties such as Aam Aadmi Party (AAP), Congress and the Shiromani Akali Dal (SAD) has been performed. The text data collected is related to the entire period of election days of electoral poll in Punjab. Therefore, the statistical technique is applied to carry out a sentiment analysis on gathered data for different political parties.

#### 3.1 *Methodology*

In this paper, the data has been collected from Twitter and pre-processing has been done on the collected text. In the proposed system, the statistical technique is used to select the political data. The statistical technique classifies the political text data into categories and further subdivided into positives or negatives.

The steps followed are:

1. First, an examination of current techniques in sentiment analysis is carried out.
2. In the second step, the data has been collected from different social media source such as Twitter.
3. Pre-processing of the data includes noise reduction, etc.
4. Various required language resources have been developed using the collected data.
5. The results have been analysed using a standard statistical technique.
6. Finally, the system is evaluated with the gathered data for different political parties.

## 4 System Architecture

The major goal is to perform the sentiment analysis for textual data collected from Twitter and stored in.txt file. After that, pre-processing of the collected data has been performed and at last, statistical techniques have been used (Fig. 1).

**Step 1:** Firstly, collected the tweets with the help of tweepy library in Python3. The tweepy library is downloaded in Python3 library by using “`pip install tweepy`”.

**Step 2:** For extracting features from these tweets, pre-processing is performed. Most of the users write comments with creative spellings, punctuations, misspellings, slang, new words, URLs and abbreviations, etc. Therefore, this type of text needs to be corrected.

### (a) Tokenization

In this step, the input is in the textual form. Tokenization is the process of breaking up text into small units. After tokenization, some unique characters such as punctuations, special symbols, etc., have been deleted. A token is a smallest unit of characters or

The screenshot shows a Visual Studio Code interface with the following details:

- File Explorer:** Shows files: `tweepy-test.py`, `preprocess.py`, `example.py`, and `dbsqlite3.py`. `tweepy-test.py` is the active file.
- Code Editor:** Displays Python code for data collection using the tweepy library. The code includes consumer keys, access tokens, and logic to search for tweets and save them to a file.
- Status Bar:** Shows "Python 3.7.3 64-bit" and other system status indicators.

```
❶ tweepy-test.py • ❷ preprocess.py • ❸ example.py • ❹ dbsqlite3.py •
❶ tweepy-test.py > ...
1 import tweepy
2 import os
3 consumerKey='XXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
4 consumerSecret='XXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
5 accessToken='XXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
6 accessSecret='XXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
7
8 authentication = tweepy.OAuthHandler(consumer_key=consumerKey, consumer_secret=consumerSecret)
9 authentication.set_access_token(accessToken, accessSecret)
10 api = tweepy.API(authentication)

11
12 numberOfTweets=input('Enter the number of tweets=>')
13 numberOfTweets=int(numberOfTweets)
14 f=open("positive.txt")
15 # for lines in f.readlines():
16 for lines in f.read().split("\n"):
17     print(lines)
18     keyword=lines
19     #keyword=input("Enter what do you want to search about?=>")
20     tweets = tweepy.Cursor(api.search, q=keyword, lang=' ').items(numberOfTweets)

21
22     for i,tweet in enumerate(tweets):
23         print((i+1),tweet.text,sep=" => ")
24         fo = open(os.path.join("tweets",keyword+".txt"),"a",encoding="utf-8")
25         fo.write(tweet.text)
26         fo.write("\n")
27         fo.close()
28
29
30
```

Python 3.7.3 64-bit ❶ ❷ ❸ ❹

**Fig. 1.** Data collection from Twitter

group of characters that makes some meaning. The unrequired symbols such as 143 punctuations, special symbols, numerical values (‘:’, ‘\*’, ‘;’, ‘:’, ‘|’, ‘/’, ‘[’, ‘]’, ‘{’, ‘}’, ‘(’, ‘)’, ‘~’, ‘&’, ‘+’, ‘?’, ‘=’, ‘<’, ‘<’, ‘|’, ‘..’, ‘?’, ‘\_’, ‘\_’, ‘\”, ‘%’, “”, ‘!’, ‘@’, ‘#’, ‘\$’, ‘%’, ‘1’, ‘2’, ‘3’, ‘4’, ‘5’, ‘6’, ‘7’, ‘8’, ‘9’, ‘0’ etc.) from the English-Punjabi code mixed data are removed during this phase.

### (b) Stop Word Removal

The variety of stop words depends upon a language. They can be used for a single multilingual stop word list. The stop words need to be eliminated as they are less relevant. To eliminate the stop words from the text data, the proposed system automatically identifies the stop words. Some common English-Punjabi stop words are as follows:

The, to, they, is, am, are, you, he, but, if, then, me, my, this, that, ਦੇ, ਦੀ, ਵਿਚ, ਦਾ, ਨੂੰ, ਹੈ, ਹੀ, ਹੋ, ਕੇ, ਉਸ, ਤੇ, ਉਹ, ਤ, ਨਾਲ, ਹੋ, ਇਹ, ਭੀ, ਨੇ, ਕਰ, ਜਿਸ, ਇਸ, ਆਪਣੇ, ਜੋ, ਮ, ਕੋਈ, ਵਾਲਾ, ਆਪ, ਤੂੰ, ਕਰਦਾ, ਕਿ, ਉਹਨਾਂ, ਜੀ, ਤਾਂ, ਕਰਨ, ਸਭ, ਜਾ, ਰਹਿੰਦਾ, ਵਾਲੇ, ਵਾਲਾ, ਹਨ, ਹੈ, ਹੋਰ, ਪਰ, ਜੇ.

**Step 3:** Finally the data is classified as positive or negative by using the proposed system. Twitter is the main source of data for analysis. The text data collected the tweets from Twitter and stored in gazetteer list.

### 1. Tweepy

Python is an interactive language. The latest version of Python3 is downloaded for windows via given link (<https://www.python.org/downloads/>). Tweepy is the leading Python library that provides access to data from twitter and it supports Python3. Tweepy is a set of Python procedures built for the purpose of sending queries to the Twitter API and providing the relevant results.

### 2. Twitter Data Collection

In this step, the procedure to create a Twitter Application Program Interface (API), to authenticate the user and creation of basic API has been discussed. It can be easily done using the URL “<https://dev.twitter.com/apps>”. After the registration process, the Consumer Key and Consumer Secret from associated application are considered. For API keys, following information is needs to be entered:

```
consumerKey = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
consumerSecret='xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
accessToken='xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
accessSecret='xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
```

### 3. Data Storage

The data has been collected from Twitter API. The English-Punjabi code mixed data has been collected from microblogging sites ranging for the period between 13 December 2017 to 31 June 2018. Around 95,800 comments have been collected

for different political parties. After cleaning the data, records have been put in the repository. The text files created comprise tweets which have been extracted using the number of keywords. “utf-8” encoding has been used for the textual data.

#### 4. Data Pre-processing

Tweets collected using the Python library are not found to be appropriate for extracting features. Generally, tweets include usernames, barspaces, special characters, stop words, emoticons, abbreviations, hashtags, time stamps, URLs, etc. So for pre-processing of the data, NLTK has been used. Removal of undesirable special symbols, hashtags, URLs, etc. and pre-processing of data has been performed by characterizing the tweets simply in terms of words so that they can be classified. A snippet of code for pre-processing of data developed in Python is as in Fig. 2.

### 5 Finding Polarity Score

The system checks the positive polarity and negative polarity in each sentence and at word level. Still, some of the words do not affect the polarity of the score like city name, person name, state name, etc. These words are matched with the BAG OF WORDS, then their polarity score is ignored. The proposed system analyses positive as well as negative polarity score using the given formula:

$$\text{POL-POSITIVE} = \sum_{i=0}^n \text{POSITIVE\_SCORE}_i$$

$$\text{POL-NEGATIVE} = \sum_{i=0}^n \text{NEGATIVE\_SCORE}_i$$

POL-POSITIVE is the positive polarity score of the sentence, POL-NEGATIVE is the negative polarity score of the sentence,  $n$  is the total number of words in the sentence,  $\text{POSITIVE\_SCORE}_i$  is the positive score of the word currently processed and  $\text{NEGATIVE\_SCORE}_i$  is the negative score of the word being currently processed.

### 6 Implementation and Results

The polarity of tweets can be expressed at different levels whether the expressed opinion in a sentence is either positive or negative. The results for opinion of people related to political tweets using the statistical techniques are as in Fig. 3.

```
File Edit Selection View Go Debug Terminal Help • preprocess.py - Data Collected  
tweepy-test.py • preprocess.py • exempl.py • dbsqlite3.py •  
Ph.D Code_Python > preprocess.py ...  
1 import re  
2 import sys  
3 from utils import write_status  
4 from nltk.stem.porter import PorterStemmer  
5  
6 def preprocess_word(word):  
7     word = word.strip('\"?!.,();')  
8     word = re.sub(r'(.)1+', r'\1\1', word)  
9     word = re.sub(r'(-|\')', ' ', word)  
10    return word  
11 def is_valid_word(word):  
12     return (re.search(r'^[a-zA-Z][a-zA-Z\\.]*$', word) is not None)  
13 def handle_emojis(tweet):  
14     tweet = re.sub(r'(:s?)|:-\)|\(\s?:|\\(-:\\\\))', ' EMO_POS ', tweet)  
15     tweet = re.sub(r'(:sD|:-D|X-?D|X-D)', ' EMO_POS ', tweet)  
16     tweet = re.sub(r'(<3|:*)', ' EMO_POS ', tweet)  
17     tweet = re.sub(r'(;?-)|;-D|(-?;) ', ' EMO_POS ', tweet)  
18     tweet = re.sub(r'(:s?)(:-|(\{\})\s?:|\\:-)', ' EMO_NEG ', tweet)  
19     tweet = re.sub(r'(:|\\|\\(|:"|\\|)', ' EMO_NEG ', tweet)  
20     return tweet  
21 def preprocess_tweet(tweet):  
22     processed_tweet = []  
23     tweet = tweet.lower()  
24     tweet = re.sub(r'((www\.(?!\.)([\S]+)|(https://)[\S]+))', ' URL ', tweet)  
25     tweet = re.sub(r'@[[\S]+]', 'USER_MENTION', tweet)  
26     tweet = re.sub(r'#([\S]+)', r'\1 ', tweet)  
27     tweet = re.sub(r'\brt\b', ' ', tweet)  
28     tweet = re.sub(r'\.{2,}', ' ', tweet)  
29     tweet = tweet.strip(' \n')  
30     tweet = handle_emojis(tweet)  
31     tweet = re.sub(r'\s+', ' ', tweet)  
32     words = tweet.split()  
33     for word in words:  
34         if is_valid_word(word):  
35             processed_tweet.append(preprocess_word(word))  
36     return processed_tweet  
37  
Python 3.7.3 64-bit 2 ▲ 1
```

**Fig. 2.** Pre-processing the data

## 7 Conclusion and Future Work

In this paper, the sentiment analysis related to the political domain has been analysed. The political data has been extracted from Twitter. Sentiment analysis of English-Punjabi code mixed data has been performed using the statistical techniques. The proposed system first tested the pipeline for unigram predictive model and accuracy is calculated. Later, the process has been repeated for 2-grams and performance has found to be marginally better as compared to the unigram model. This work can be extended in future by checking the sentiments of emoticons. Identifying the emoticons and replacing them with a single word are challenging task ahead.

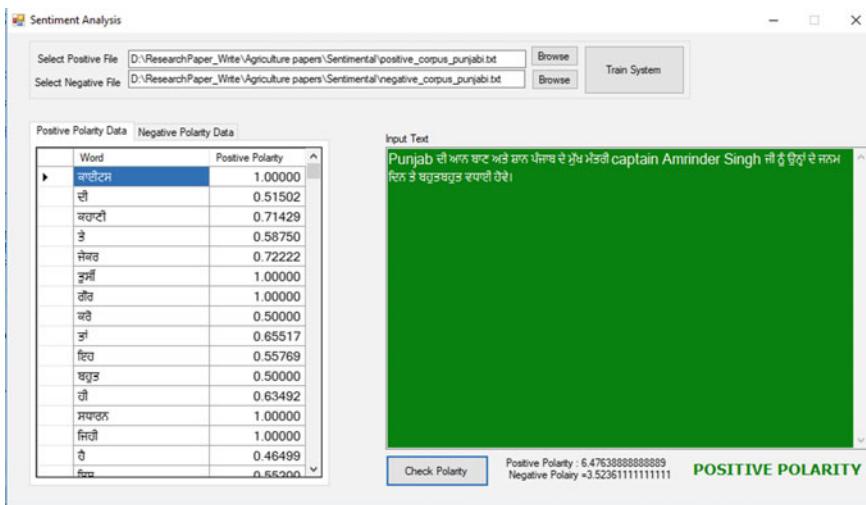


Fig. 3. Result of sentiment analysis

## References

- Kušen E, Strembeck M (2018) Politics, sentiments, and misinformation: an analysis of the Twitter discussion on the 2016 Austrian presidential elections. *Online Soc Netw Media* 5:37–50
- Jain EU, Sandhu A (2015) Emotion detection from Punjabi Text using hybrid support vector Machine and maximum entropy algorithm. *Int J Adv Res Comput Commun Eng (IJARCCE)* 4(11):89–93
- Pooja P, Sharvari G (2015) A survey of sentiment classification techniques used for Indian regional languages. *IJCSA* 5(2):13–26
- Cambria E (2016) Affective computing and sentiment analysis. *IEEE Intell Syst* 31(2):102–107
- Laryea BNL et al (2015) Web application for sentiment analysis using supervised machine learning. *Int J Softw Eng Appl* 9(1):191–200
- Wong FMF et al (2016) Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Trans Knowl Data Eng* 28(8):2158–2172
- Boutet A, Kim H, Yoneki E (2012) What's in your tweets? I know who you supported in the UK 2010 general election. In: Sixth international AAAI conference on weblogs and Social Media
- Golbeck J, Hansen D (2011) Computing political preference among twitter followers. In: Proceedings of the SIGCHI conference on human factors in Computing systems. ACM
- Pennacchiotti M, Popescu A-M (2011) Democrats, republicans and starbucks aficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM
- Kaur S, Rashid EM (2016) Web news mining using Back Propagation Neural Network and clustering using K-Means algorithm in big data. *Indian J Sci Technol* 9(41)
- Rashid M, Hamid A, Parah SA ?(2010) Analysis of streaming data using big data and hybrid machine learning approach. In: *Handbook of multimedia information security: techniques and applications*. Springer, Cham, pp 629–643
- Saif H, He Y, Alani H (2012) Semantic sentiment analysis of twitter. In: *International semantic web conference*. Springer, Berlin, Heidelberg

13. Ramanathan V, Meyyappan T (2019) Twitter Text mining for sentiment analysis on people's feedback about Oman tourism. In: 2019 4th MEC international conference on big data and smart city (ICBDSC). IEEE
14. Terán L, Mancera J (2019) Dynamic profiles using sentiment analysis and twitter data for voting advice applications. Gov Inf Q
15. Abid F et al (2019) Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. Future Gener Comput Syst 95:292–308

# Automatic Understanding of Code Mixed Social Media Text: A State of the Art



Neetika, Vishal Goyal, and Simpel Rani

**Abstract** Social media content is often addressed as noisy or informal text due to the existence of zigzag conversational patterns. People do not always use Unicode rather they mix multiple languages. Hence, the processing of code mixed data postures computational challenges ahead. Since decades, social media content and its analysis have gained momentum worldwide. In parallel, the pace of research on Indian languages is also commendable. In India, the users of social media hail from different religions, regions, subdivisions and culture. The major concern of the paper is to throw light on the works done in Indian languages with code mixed social media as a concern. The journey of the research in the respective field has various milestones between basic tasks of natural language processing and deep learning. This paper focusses on the works done on Indian languages with respect to language identification, normalization and POS tagging. Efforts have been done to discuss the tools, techniques and the corpora used by researchers in different Indian languages. In the digital age, we have an abundance of tools and APIs available for extracting code mixed text. Still, there is paucity of public data available for analysis. The need of the hour seems to be protruding toward deep learning and extending the public availability of code mixed corpora.

**Keywords** Code mixing · Language identification · Normalization · Part of speech (POS) tagging

---

Neetika (✉)

Department of Computer Science, College of Engineering & Management, Rampura Phul, Punjab, India

e-mail: [sunshine\\_neetika@yahoo.com](mailto:sunshine_neetika@yahoo.com)

V. Goyal

Department of Computer Science, Punjabi University, Patiala, Punjab, India

e-mail: [vishal.pup@gmail.com](mailto:vishal.pup@gmail.com)

S. Rani

Department of Computer Science and Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

e-mail: [simpel\\_jindal@rediffmail.com](mailto:simpel_jindal@rediffmail.com)

## 1 Introduction

The revolutionary induce of social media has given a new perspective to cultural enhancements, as people of different cultures can communicate with one another in a formal as well as informal manner. Worldwide English is still by far the most popular language in SMC, though its dominance is receding. Hong et al. [1] analyzed that only half of the tweets were in English within the dataset of 62 million tweets. They developed an automatic language detection algorithm to identify the top ten popular languages on Twitter.

Automatic understanding of social media text becomes a difficult task when a bilingual/multilingual speaker frequently switches between languages. There are variations in writing styles of users, and due to autofill feature the complexity of analysis of code mixed text also increases. Researchers have stated number of reasons regarding ‘Why do people Code Mix?’ The basic reasons sorted out so far are for role identification, style identification, choice of topic, inability of expressions, impact and effective speech, emotional arousal, ethnic identity, showing off or showing kinship.

In some cases, it is done deliberately to exclude a person from a conversation. In other cases, the topic decides the language, so the utterances change with the demand of the topic. The phenomenon of code mixing seems to be prevalent with every peck of a second; a person communicates using social media. The users of social media keep typing at a fast pace or due to autofill feature in devices which increases typo errors. Some words commonly misspell like media (as mcdia), accept (as acept). Most of the users on social media creatively use emoticons, meta tags, URL tags and hash tags. Since decades, code mixing has gained existence in movies, songs, advertisements, etc. The main focus of the advertising companies and commercials is to gain the attention of people of that particular area. Examples: “Carry on jatta”; Coca Cola, “Thanda matlab coca cola”. Often names of persons match with movies names, places or objects and such words need to be handled as special cases. Example: “Everyone is excited when Apple will launch a new iphone?” With context to Indian languages, there is one interesting challenge. Some reduplicated words like dance-vance, super-duper, etc., are commonly used.

Jauhainen et al. [2] have discussed the methods and features used for language identification by using a unified notation. In addition, applications and evaluation methods of LID are also discussed. All the pertaining issues, complete survey of the work done, have been mentioned in the paper.

## 2 Evolution of Code Mixed Social Media Text

Since decades, code mixed social media text has been coined with different names like anglicism, borrowing, code alternation, code mixing and code switching, etc.

## ***2.1 Bilingualism/Multilingualism***

Social media users strongly contribute to bilingual or multilingual in using social media content. India comprises social media users from various regions, subdivisions, religions, castes, etc.

## ***2.2 Borrowing***

Borrowing means inserting a foreign language word into native language. It is strongly agreed that there exists a sort of continuum between code mixing and loan vocabulary. Many linguists believe that loan words start out as code mixing or nonce borrowing but when used repeatedly across the languages they automatically become a part of the native vocabulary and acquire the characteristics of the “borrowing” language. Lots of cases have been triggered when there is no clear distinction between borrowing, anglicism, code mixing or code switching.

## ***2.3 Code Switching***

Code switching occurs when one starts writing in one language and then switches to another one. Code switching is divided into four categories: intra-sentential code switching, inter-sentential code switching, intra-word code switching and tag switching or extra sentential code switching. Intra-sentential code switching occurs within a clause or sentence boundary, within a clause or sentence, with no interruptions, hesitations or pauses. Inter-sentential switching occurs at a clause or sentence boundary, where each clause or sentence is in one language or the other. Intra-word code switching occurs when change is within a word boundary. Tag switching or extra sentential code switching occurs when certain set of phrases in one language are inserted into an utterance from another language, similar to intra-code sentential switching.

## ***2.4 Code Mixing***

Code mixing means using two languages in an utterance or mixing words of two languages. Code mixing is divided into three categories: intra-utterance code mixing, inter-utterance code mixing, and word level code mixing. Intra-utterance code mixing takes place in a single sentence or utterance as the speaker is proficient in both languages. Example: “Dr. Ambedkar ne apni book “Problem of Rupee” mein likha tha ki bhrishtachar khatam karne ke liye 10 saalon mein note badal dene chahiye”.

Inter-utterance code mixing occurs when one changes from one language to another in the same conversation/utterance. Example: “main apne parivar ke saath hun and mi husband”. Word level code mixing is the smallest unit of code mixing. It captures intra-word code mixing and includes cases where code mixing takes place within a single word. Example: “he is a bigdefied child of his parents” (bigde (hindi) + fied (English suffix).

### **3 Basic Tasks Involved in Automatic Understanding of Code Mixed Social Media Text**

Cetinoglu et al. [3] have addressed various computational challenges of natural language processing (NLP) on code mixed data. Natural language processing involves some basic tasks in understanding code mixed social media text. They are normalization, language identification, language modeling, part of speech tagging, parsing, machine translation and automatic speech recognition. Also, these tasks are many times interrelated, and textual information serves as input for one another.

#### ***3.1 Normalization***

Text normalization is the task of standardizing text that deviates from some agreed-upon (or canonical) form. Usually, social media text is written using romanized typing. The challenges posed for normalization bear mapping back to respective language used.

#### ***3.2 Language Modeling***

In language modeling, the probabilities are assigned to the text and utterances. Models such as n-gram models, factored language models and neural language models are used in many NLP applications such as machine translation and automatic speech recognition. Tokenization and normalization are the initiating steps to prepare the training data.

#### ***3.3 Language Identification***

Language identification (LID) is considered as one of the successfully accomplished tasks among computational code switching approaches. Language identification is

an essential prerequisite for automatic text processing. Using dictionary approaches, it can be considered as a solved problem for monolingual text in which n-gram approaches, word embeddings and stop word lists can reach up to 100% accuracy.

### ***3.4 POS Tagging***

Next to language identification, POS tagging is another indispensable area of research. POS tagging assigns tag from different tagset categories to every individual token. Similar to parsing, it also considered an important prerequisite for processing text. For monolingual data, the POS taggers are available for most of the languages but not for code mixed datasets.

### ***3.5 Parsing***

Parsing is the task of determining syntactic relations between words and phrases of a given sentence. Building a bilingual tagger and parser has been considered as a strong challenge for parsing code mixed data. The challenge becomes tougher when two language systems are contextually similar.

### ***3.6 Machine Translation***

With help of machine translation (MT), it has become easier to translate text among different languages. Translation systems for mixed text are not yet available due to paucity of data. Machine translation faces challenges for dual systems lexically as well as syntactically.

### ***3.7 Automatic Speech Recognition***

Automatic speech recognition (ASR) system transforms speech signal to text as an essential component. There are a number of phonetic tools like SOUNDEX and Metaphone are used for the system. Automatic speech recognition (ASR) system transforms speech signal to text as an essential component.

## 4 A Brief Review of Works Done with Context to Indian Languages

Works done on Indian languages with respect to language identification, normalization and POS tagging are mainly focussed on the paper. Efforts are done to include the type of datasets, tools and techniques used by various researchers on Indian languages code mixed corpora.

Researchers have mostly used unsupervised dictionary method or supervised learning models and classifiers such as Naive Bayes (NB), n-grams approach, support vector machines (SVM), conditional random fields (CRF), hidden Markov model (HMM), logistic regression (LR), long short-term memory (LSTM), bidirectional LSTM and recursive neural networks (RNN). The journey of the research in the respective field has various milestones between basic tasks of natural language processing and deep learning.

Besides basic tasks of natural language processing (NLP) like language identification, named entity recognition (NER), part of speech (POS) tagging lot of research are being carried out on sentiment analysis [4], question answering, etc. Furthermore, some of the emerging research patterns are hate detection [5], stance detection, dialect identification, cyberbullying detection, humor detection and aggression detection.

Thara and Poornachandran [6] have also discussed applications of code mixed data. In addition, different algorithms, problems and challenges faced by researchers in respective field are also shared. Code mixed text has become important for policymaking as well. A multilingual text corpus can be used with diversity. Normalized text can act as a strong input for sentiment analysis. By removing the noise from the noisy data, better input can be provided to the text reading software for a best output to the blind. Normalized text can be used for information retrieval. Students' choices, interests and skills can be well understood through code mixed chats.

Social media interaction and understanding can be enhanced with diversity of languages. It has become beneficial for marketing industry as they can come to know people choices, interest through reviews. Travel agencies can also gather information regarding tastes and preferences of people through posts and reviews. Multilingual books can be published. In online game, interaction between persons of various countries can be increased worldwide. Law and order can be maintained by tracking and analyzing gangsters', criminals' posts, etc. Language identification and normalized text can be used as input for text-to-speech (TTS) system. During disasters, multilingual problems and data can be exchanged and analyzed. Duarte et al. [7] have discussed in detail the strengths and weaknesses of text classifiers. They have tried to communicate to the policymakers the limitations of tools used by researchers. They have tried to throw light on the strengths of the text classifiers. Also, the main aim highlighted is to communicate the findings to the policymakers.

Pavan et al. [8] proposed a Romanized text language identification system (RoLI). They experimented on five Indian languages: Hindi, Telugu, Tamil, Kannada and Malayalam for sound-based similarity of words with 98.3% accuracy. Sequiera et al. [9] suggested an algorithm for back transliteration of words into Kannada script. Word

level language identification and normalization were performed on Bengali–English, Gujrati–English, Kannada–English, Malayalam–English and Tamil–English word pairs. Bali et al. [10] analyzed English–Hindi Facebook comments and found that embedding of Hindi words in English follows some formulaic patterns of nouns and particles. Das and Gamback [11] performed word level language identification on English–Bengali and English–Hindi Facebook messages by adopting n-gram language profiling and pruning, dictionary-based detection and SVMs.

Gamback and Das [12] performed language identification on English–Bengali, English–Hindi dataset [13] by adopting n-gram language profiling and pruning, dictionary-based detection and support vector machines (SVM). The accuracy reported is 96% for English–Bengali and 98% for English–Hindi using SVM. Gokul-Chittaranjan et al. [14] performed word level language identification with CRFs using lexical and character-based features and max entropy classifier. CRFs outscored 95% accuracy for English–Spanish, English–Mandarin and English–Nepali data and 85% for standard Arabic–Arabic data. Vyas et al. [15] adopted the initial work on part of speech (POS) tagging of English–Hindi code mixed social media content. It was deduced that POS tagging performs well with language identification and transliterated form of words.

Das and Gamback [16] introduced the code mixing index for computing the level of code mixing in the code mixed corpora developed by [13] blended with Facebook data. Kaur and Singh [17] experimented rule-based approach on normalization of 1000 English–Punjabi sentences. The precision value of 52.3 was achieved. Desai and Narvekar [18] suggested a word shortening algorithm to handle noisy data which proved better than traditional models. Sequiera et al. [19] performed language identification and POS tagging using their own model VGSBC using n-gram identifiers. In addition, they reported an accuracy of 84%. Jamatia et al. [20] experimented POS tagging, code mixing index (CMI) using CRF, sequential minimal optimization, Naive Bayes and random forests on English–Hindi Twitter and Facebook data. They adopted CMU tokenizer, Google’s universal tagset [21], Twitter-specific tagset [22] and Indian languages POS tagger [23]. The accuracy reported was 87%. Similar approach was used by [24] for Gujarati–English corpora.

Dutta et al. [25] experimented on code mixed English–Bangla text by using noisy channel model and CRF++. Word level language identification accuracy reported was 90.5% and spell checker accuracy as 69.43%. Sharma and Motlani [26] worked on POS tagging of Hindi–English, Bengali–English and Tamil–English datasets. The HMM outscored with 80.68% accuracy for Bengali–English and Tamil–English. Sitaram et al. [27] experimented Hindi–English, German–English data for language identification, normalization and synthesis. They adopted Naive dictionary model, HMM and max entropy using Viterbi decoding. HMM was best for Hindi–English with accuracy of 89% and for German–English with accuracy of 96%.

Sharma et al. [28] used a pipeline for Hindi–English code mixed social media text dataset. Language identification, normalization, POS tagging and shallow parsing were done using normalized frequency of the word in British National Corpus, LEXNORM; a binary feature indicating presence of a word in the lexical normalization dataset [29], HINDI DICT [30]; a dictionary of 30,823 Hindi transliterated

words, n-gram approach, noisy channel Giza++ [31], SILPA Libindic spell checker, SVM and prefixes and suffixes of the words. They reported normalizer average accuracy as 74.48% and pipeline accuracy as 75.07%. Ranjan et al. [32] experimented on Tamil–English and Telugu–English twitter messages. They adopted Weka2 (Naive Bayes), SVM and random forests for training the dataset. In addition, they also adopted recurrent neural network language modeling toolkit (RNNLM) [33]. They reported random forest as best for Tamil with 84.48% accuracy and for Telugu with 86.02% accuracy. Using perplexity measures, it was deduced that the code mixed data of Indian social media has very less similarity to the normal data.

Phadte et al. [34] performed word level language identification on Konkani–English dataset. They developed wordlists for English–Konkani (15,195 words) language pairs. 97% accuracy was reported using CRFs. Veena et al. [35] reported accuracy of 95% for Tamil–English and 93% for Malayalam–English data. Lakshmi and Shambhavi [36] worked on Kannada–English code mixed corpora by blending dictionary module, classifier, bag-of-words (BOW) and term frequency and inverse document frequency (TF\_IDF). Jamatia et al. [37] performed deep learning-based language identification in English–Hindi, English–Bengali, English–Hindi–Bengali code mixed data. Accuracy reported was 88.27 adopting LSTM for English–Bengali text, 86.97 adopting LSTM for English–Hindi as and 88.27 adopting Bi-LSTM for English–Hindi–Bengali. Gupta et al. [38] experimented on dataset of ICON\_2016 with CRF classifier for POS tagging. Jamatia et al. [39] worked on annotation, CMI, utterance boundary detection and POS tagging on English–Bengali, English–Hindi tweets and English–Hindi Facebook messages. Utterance boundary detection was found as challenging, and POS tagging has to be based on context [40].

## 5 Conclusion

Since decades, the research on analysis of code mixed social media text has gained attention as area of research especially in India. Deep learning has penetrated into the scenario contributing to higher accuracy levels. Furthermore, researchers have introduced recursive neural network (RNN) techniques like long short-term memory (LSTM), bidirectional LSTM heading to convolutional neural networks. In addition, handling the scarcity of data, the LSTM approach also improves on the word embeddings where words similar in meaning or used in similar context are learnt using deep learning. Need of the hour is to follow up cross-lingual transfer of models. Still there are challenges ahead. They can be resolved by handling more deep learning models like convolutional neural networks. The minority languages which are under diminishing phase can gain momentum with active participation of new researchers in the research area. Efforts have been done to discuss the tools, the techniques and the corpora used by researchers in different Indian languages. In the digital age, we have the abundance of tools and APIs available for extracting code mixed text. Still there is paucity of public data available for analysis. The need of the hour seems to be protruding toward deep learning and extending the availability of code mixed corpora.

## References

1. Hong L, Convertino G, Chi EH (2011) Language matters in twitter: A large scale study. In: Fifth international AAAI conference on weblogs and social media
2. Jauhainen TS, Lui M, Zampieri M, Baldwin T, Lindén K (2019) Automatic language identification in texts: a survey. *J Artif Intell Res* 65:675–782
3. Cetinoglu O, Schulz S, Vu NT (2016) Challenges of computational processing of code-switching. In: Second workshop on computational approaches to code switching, pp 1–11
4. Konate A, Du R (2018) Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques. *Wuhan Univ J Nat Sci* 23(3):237–243
5. Santosh T, Aravind K (2019) Hate speech detection in Hindi-English code-Mixed social media text. In: Proceedings of the ACM India joint international conference on Data Science and Management of data. ACM, pp 310–313
6. Thara S, Poornachandran P (2018) Code-mixing: a brief survey. In: International conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 2382–2388. <https://doi.org/10.1109/icacci.2018.8554413>
7. Duarte N, Llanso E, Loup A (2018) Mixed messages? The limits of automated social media content analysis. *FAT*, p 106
8. Pavan K, Tandon N, Varma V (2010) Addressing challenges in automatic language identification of romanized text. In: 8th International conference on natural language processing (ICON-2010)
9. Sequiera, R. D., Rao, S. S., & Shambavi, B. R.: Word-Level language identification and back transliteration of Romanized text: a shared task report by BMSCE. In: MSRI FIRE working notes. (2014)
10. Bali K, Sharma J, Choudhury M, Vyas Y (2014) I am borrowing ya mixing? An analysis of English-Hindi Code mixing in Facebook. In: Proceedings of the first workshop on computational approaches to code switching, pp 116–126
11. Das A, Gamback B (2015) Code-mixing in social media text: the last language identification frontier? *Revue TAL* 54(3):41–64
12. Gamback B, Das A (2014) On measuring the complexity of code-mixing. In: 11th international conference on natural language processing, pp 1–7, Goa
13. Barman U, Das A, Wagner J, Foster J (2014) Code mixing: a challenge for language identification in the language of social media. In: First workshop on computational approaches to code switching, pp 13–23
14. GokulChittaranjan, Vyas Y, Bali K, Choudhury M (2014) A framework to label code-mixed sentences in social media. In: First workshop on computational approaches to code-switching. ACL, Doha
15. Vyas Y, Gella S, Sharma J, Bali K, Choudhury M (2014) Pos tagging of english-hindi code-mixed social media content. In: Conference on empirical methods in natural language processing (EMNLP), pp 974–979
16. Das A, Gamback B (2014) Identifying languages at the word level in code-mixed indian social media text. In: 11th International conference on natural language processing, pp 378–387
17. Kaur J, Singh J (2015) Toward normalizing romanized gurumukhi text from social media. *Indian J Sci Technol* 8(27):1–6
18. Desai N, Narvekar M (2015) Normalization of noisy text data. *Procedia Comput Sci* 45:127–132
19. Sequiera R, Choudhury M, Bali K (2015) Pos tagging of Hindi-English code mixed text from social media: some machine learning experiments. In: 12th international conference on natural language processing, pp 237–246
20. Jamatia A, Gamback B, Das A (2015) Part-of-speech tagging for code-mixed English-Hindi twitter and facebook chat messages. In: International conference recent advances in natural language processing, pp 239–248
21. Petrov S, Das D, McDonald R (2012) A universal part-of-speech tagset. In: Eighth international conference on language re-sources and evaluation (LREC-2012). European Languages Resources Association (ELRA), Turkey, pp 2089–2096

22. Gimpel K, Schneider N, O'Connor B, Das D, Mills D (2010) Part-of-speech tagging for twitter: annotation, features, and experiments. Technical Report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science
23. Baskaran S, Bali K, Bhattacharya T, Bhattacharya P, Jha GN (2008) A common parts-of-speech tagset framework for indian languages. In: LREC 2008
24. Dholakia PS, Yoonus MM (2014) Rule based approach for the transition of tagsets to build the POS annotated corpus. *Int J Adv Res Comput Eng* 3(7):7417–7422
25. Dutta S, Saha T, Banerjee S, Naskar SK (2015) Text normalization in code-mixed social media text. In: 2nd international conference on recent trends in information systems (ReTIS). IEEE Press, New York, pp 378–382
26. Sharma A, Motlani R (2015) Pos tagging for code-mixed indian social media text: systems from IIIT-h for icon NLP tools contest
27. Sitaram S, Rallabandi SK, Rijhwani S, Black AW (2016) Experiments with cross-lingual systems for synthesis of code-mixed text. In: SSW, pp 76–81
28. Sharma A, Gupta S, Motlani R, Bansal P, Shrivastava M (2016) Shallow parsing pipeline for Hindi-English code-mixed social media text. In: NAACL-HLT, pp 1340–1345
29. Han B, Baldwin T (2011) Lexical normalisation of short text messages: Makn sens a# twitter. In: 49th Annual meeting of the Association for Computational Linguistics: Human Language Technologies, pp 368–378
30. Gupta K, Choudhury M, Bali K (2012) Mining Hindi-English transliteration pairs from online Hindi lyrics. In: LREC, pp 2459–2465 (2012)
31. Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. *Comput Linguist* 29(1):19–51
32. Ranjan P, Raja B, Priyadarshini R, Balabantary RC (2016) A comparative study on code-mixed data of Indian social media vs formal text. In: 2nd international conference on contemporary computing and informatics (IC3I). IEEE, pp 608–611
33. Mikolov T, Kombrink S, Deoras A, Burget L, Cernocky J (2011) RNNLM-recurrent neural network language modeling toolkit. In: ASRU Workshop, 2011, pp 196-201
34. Phadte A, Wagh R (2017) Word level language identification system for Konkani-English code-mixed social media text (CMST). In: 10th annual ACM India compute conference. ACM, pp 103–107)
35. Veena PV, Kumar MA, Soman KP (2017) An effective way of word-level language identification for code-mixed face-book comments using word-embedding via character-embedding. In: International conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 1552–1556
36. Lakshmi BS, Shambhavi BR (2017) An automatic language identification system for code-mixed English-Kannada social media text. In: 2nd international conference on computational systems and information technology for sustainable solution (CSITSS). IEEE Press, pp 1–5
37. Jamatia A, Das A, Gamback B (2019) Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora. *J Intell Syst* 38(3):399–408
38. Gupta D, Tripathi S, Ekbal A, Bhattacharyya, P (2017) SMPOST: parts of speech tagger for code-mixed indic social media text. arXiv preprint [arXiv:1702.00167](https://arxiv.org/abs/1702.00167)
39. Jamatia A, Gamback B, Das A (2016) Collecting and annotating Indian social media code-mixed corpora. In: International conference on intelligent text processing and computational linguistics. Springer, pp 406–417
40. Mave D, Maharjan S, Solorio T (2018) Language identification and analysis of code-switched social media text. In: Proceedings of the third workshop on computational approaches to linguistic code-switching, pp 51–61. <https://doi.org/10.18653/v1/w18-3206>

# Secure Server Virtualization Using Object Level Permission Model



Varsha Grover and Gagandeep

**Abstract** Virtualization is a framework or methodology that is used to divide the resources of a computer into multiple execution environments by applying concepts or technologies storage virtualization, client virtualization, and server virtualization are different ways to achieve virtualization. Our goal in this paper is to implement server virtualization with attacks and threats with possible solutions. Server virtualization is best suitable for dividing a server into a multiple virtual servers via software such as VMware and Hyper V. This method is used to increase the resource utilization by performing the same. In this each virtual server act and look like a physical server by increasing the capacity of every single physical machine (Kuche et al. in Int J Comput Netw Wireless Mobile Commun (IJCNWMC) 4:5–10 (2014), [1] with problems like efficient resource management, more downtime and privacy and security of data are associated with server virtualization. The problem is unauthorized users are accessing the access of machine/server in VMware platform. This can be done by proposing model named **object level permission model** to provide security by assigning permissions and roles as per level.

**Keywords** Server virtualization · Security · Attacks and threats · VMware

## 1 Introduction

Cloud computing is developed to enable the information technology world for utilizing computer resources efficiently and more proficiently. Cloud computing is an efficient way to increase the capacity, dynamic scalability or add capabilities using virtualization resources, platform, infrastructure, and software as service that can be accessed over the internet. Virtual machines (VMs) play an important role to improve the utilization of cloud resources. VMs are a virtual computer similar to a physical

---

V. Grover (✉) · Gagandeep  
Department of Computer Science, Punjabi University, Patiala, India  
e-mail: [grover.varsha5@gmail.com](mailto:grover.varsha5@gmail.com)

computer in which application or operating system are installed and run. Virtualization is an innovative technology, which is significantly expanding in the information technology industry. It provides multiple logical resources on a single server. Server virtualization is a virtualization that helps to divide a server into many virtual servers with the help of software. This method is useful for increasing resource utilization by dividing physical servers into multiple servers, so that every small machine behaves and acts like a server [2].

**For instance**, when a small company develops a website, instance of purchasing a physical server, server virtualization technique can be performed with the help of shared hosting that helps to reduce cost and efficient management. This paper identifies the problem of attacks and threats which occur while implementing server virtualization in cloud computing. It also experiences slow speed since other users request the same resources. There are various attacks and threats that are associated with server virtualization such as VM sprawl, hijacking through the self-service portal and Cloud service provider API risks and various solutions are available in VMware ESXi. There are inbuilt features like CPU isolation, memory isolation, and device isolation and lockdown mode. Some features like Certificate replacement and Smart card authentication methods can be configured in this model. In this paper, all these methods will be explored and accordingly model will be proposed to solve these issues.

## 2 Server Virtualization

Server virtualization is used to divide server into multiple virtual servers. There are number of server resources that include servers, processors, and operating systems. A software is used to divide server into multiple virtual servers by administrator. This helps to increase resource utilization and CPU utilization by increasing capability [3, 4].

To implement server virtualization, Microsoft hyper-v and VMware ESXI tools are used. Microsoft hyer-v comes with few limitation such as compatibility with Microsoft window(O S). On the other hand, VMware ESXI works with all operating system. VMware EXSI 4.0 to 6.0 are versions available. To implement VMware EXSI 6.0, machine requires minimum 10 GB RAM, 160 GB hard disk and 4 processors and Click to <https://www.vmware.com/products/vsphere-hypervisor.html> link to download VMware EXSI.

### Following are the steps need to be followed

- Make account to VMware by register
- After login, download the VMware EXSI 6.0
- Run the VM Machine

Machine is ready to use.

Various parameters are associated with server virtualization like memory, CPU, hard disk, network adapter, and display. Customization of these parameters can be done by right click on VMware EXI 6.0 and setting option. New hardware can be added and existing can be removed. Run the VMware EXI 6.0. IP Address is generated 192.168.1.5 after running VMware EXI 6.0. This IP address is accessed in web browser. Click on Advance button and then proceed to this IP. Fill username and password that is created at the time of VMware EXSI installation. VMware EXSI is ready to use.

## ***2.1 Threats and Attacks on VMware EXSI***

Security threats and attacks can be broadly categorized as either internal threats that originate within the system and its tenants, or external threats that originate outside the system. This latter category includes threats to the infrastructure created to host a server group. Virtual environments have many security attacks and risks as physical environments, and it takes different precautions to secure these virtual servers. Followings are the threats and attacks of server virtualization [5–7].

### **2.1.1 Multi-tenancy and Internal Threats**

An Internal threats are threats that occurs when an individual or a group of any organization seeks to exploit organizational asset. Direct visibility or access to most system level resources that includes information of physical host like IP addresses, MAC addresses, CPU type, ESXi access, physical storage locations are not provided in vCloud Director. Users attempt to access to information about the infrastructure of system where the cloud-enabled applications are running. Therefore, if user is able to perform the same there is always a scope of having attack at the lower level. Also at the virtualization level, users try to access the unauthorized control to resources of another organization. Users try to take privilege that is kept for administrators only. Users perform some actions that may or may not, disrupt the availability and performance of the system. A “denial of service” occurs in special cases for some users.

### **2.1.2 Secure Hosting and External Threats**

External threats refer to those threats which occur from outside the cloud such as attacking vCloud Director through APIs and Web interfaces. A unauthorized user

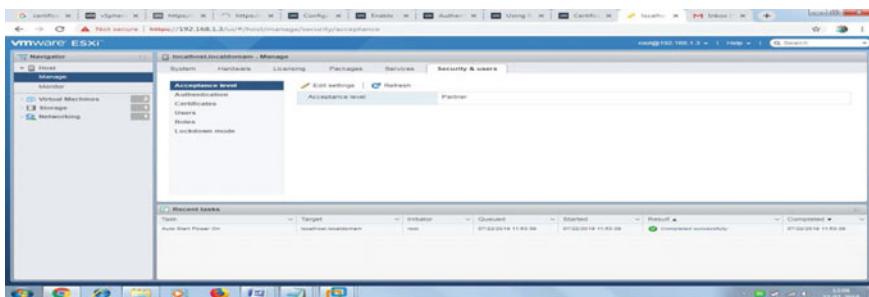
who has no access rights to the system can also gain access as an authorized user when this threat occurs. Another sources of external threats like authenticated users exploit vulnerabilities that take access to the system. Roles of users try to attempt the loop holes in the system implementation in order to get the information, acquire access to services of the cloud through unavailability of system. Some of the attacks try to disrupt the tenant boundaries and hardware abstraction layers. The deployment of the variety of layers of the system affects the mitigation of these threats like the externally facing interfaces by including firewalls, routers, VPNs, and so on.

## 2.2 Securing VMware EXSI

ESXi hypervisor architecture has many inbuilt security features like CPU isolation, memory isolation, and device isolation. Lockdown mode, certificate replacement, and smart card authentication are advance methods that are used to increase security. A firewall is protected with EXSi [8, 9] (Fig. 1).

### 2.2.1 Certificate Replacement

Certificate replacement is one of the method used in VMware EXSi to provide security. Certificates are replaced in this method that depends upon the company policy and requirements. CLIs with installation are used for configuring certificate replacement. VMCA (VMare Certificate Authority) is part of Platform Services Controller and embedded deployment in which each node, each vCenter user, and each ESXi host takes provision by VMCA with a certificate.



**Fig. 1** Security methods of VMware EXSI

### 2.2.2 Lockdown Mode

Lockdown method is best method for increasing the security of hosts, only vpxuser has authentication to perform operations against the host directly in this mode. All operations are performed through vCenter Server. There are two types of Lockdown Mode that exist—Normal Lockdown mode or Strict Lockdown mode.

**To enable or disable lockdown mode from the DCUI:**

Log into the ESXi host.  
Click on **Manage**  
Press **Security and users in tab**  
Press **Lockdown Mode** and it will be enabled

### 2.2.3 Smart Card Authentication

Smart card authentication is best method that is used in security to identify the real user. Many government agencies and large enterprises use smart card for authentication to increase the security of systems. These cards are plastic cards which are embedded with integrated circuit chip. Users connect their smart card and software on the host computer to interact with the keys material. Cryptographic keys and other secrets stored on the card keep protected the data in both ways physically and logically. Smart card authentication use to log into the ESXi by using a Personal Identity Verification and Common Access Card.

**To enable Smart Card Authentication**

Log into the ESXi host.  
Click on **Manage**  
Press **Security and users in tab**  
Press **Authentication and Smart Card Authentication Enabled**.  
Click on **Join Domain**

## 3 Comparison Analysis of Security Techniques

Every threats and attacks have its own effects on server. Various techniques can be used to get rid of attacks related to server virtualization. In this paper, various

**Table 1** Comparison of security techniques

Technique	Features
Lockdown method	It provides authentication to perform operations against the host
Certificate replacement	Certificates are used for communication between vsphere components
Smart card authentication	To identify the real users by physical identification
Assigning permissions using Object level model	To give permission to each user as per their role and level

techniques of server virtualization with its comparison will be carried out to find out the best technique among all (Table 1).

## 4 Experimental Setup and Results

To overcome the various threats and attacks, **Object level permission model** proposed in this paper. This model works using vSphere object hierarchy by giving permissions to the objects. For example, user has a read only role on one VM and the admin role on another VM at the same time. Followings are the components of this model [10, 11].

### 4.1 Permissions

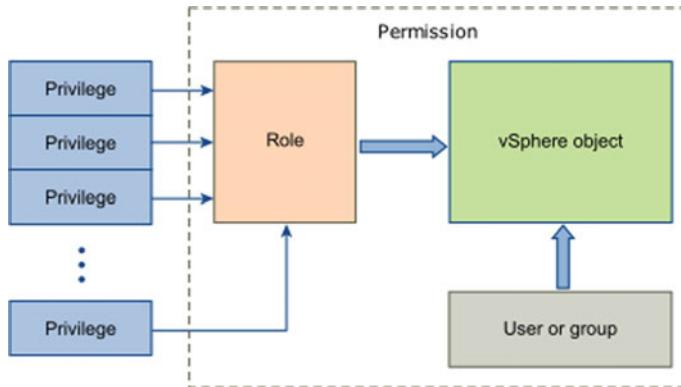
Permission refers to privilege given to users. Each object is associated with permissions in the hierarchy.

### 4.2 Users and Groups

Users or Groups, who are belonging to authenticated users to whom privileges can be assigned.

### 4.3 Privileges

Privileges are the access controls given to users depends upon their role. Those privileges can be grouped into roles as per user.



**Fig. 2** Object level permission model

#### 4.4 Roles

Roles play an important part for assigning the permissions depending upon the types of users. Some roles cannot be changed like administrator (Fig. 2).

In this paper, we implemented Object level permission model with VMware EXSi as a tool. This method works in the way in which we assign role to each user by assigning permissions. There are different roles of every user depending upon the access right that we want to permit. The users can be categorized as root and anonymous. The different types of roles which are provided to the user are administrator, no access, no cryptography administrator, read only, and view.

**To add a role the following step is required:**

**Click Home > Administration > Roles > Add Role.**

After a role is added then different types of privileges are associated. For data store, there is allocation of space privilege and browse data privilege. For host, there are different privileges associated like for local operations, for creating virtual machine, deleting virtual machine, and managing user groups and reconfiguring virtual machine. There is also network assignment privilege. Other privileges include assigning virtual machine to the resource pool.

**By following these steps we can add the permission at the highest level and set to propagate the permissions**

Right-click the Inventory object, then click the **Permissions** tab.

Click **Add Permissions**.

Click **Add** and in the **Domain** field, select the Active directory.

Select the role from the Assigned Role dropdown.

Deselect **Propagate to Child Objects**.

## 5 Conclusion

This paper reports an empirical study on the security in context of server virtualization. Various security threats and issues like data protections, resource utilizations, CPU response time are studied in this paper with possible solutions and methods. To overcome, these threats and attacks various security methods are proposed like lock-down mode, certificate replacement, and smart card authentication and proposing a model named as Object level permission model. Every method has its own features to provide security. Customization of these methods can be done by applying respective settings. This paper proposes Object level permission model which is implemented to assign the permissions depending upon the role of users so that no unauthorized user can take the access of server. Security can be enhanced and overcome the problem of external threats by assigning the privilege to user as per its role. Although said model performs well with some issues like unauthorized users can access of the server by cracking password using few commands that need to be resolved in future with smart card authorization method and it is one of the emerging concept, that will be discussed in future.

## References

1. Kuche AJ, Dakhne DM, Pardi RL (2014) Towards future smart phone: mobile virtualization using smart phone. *Int J Comput Netw Wireless Mobile Commun (IJCNWMC)* 4:5–10
2. Mod P, Bhatt M (2014) A survey on dynamic resource allocation technique in cloud environment—a survey. *Int J Adv Res Comput Eng Technol* 3:7815–7818
3. Durairaj M, Kannan P (2014) A study in virtualization techniques and challenges in cloud computing. *Int J Sci Technol Res* 3:147–151
4. Prohit B, Sharma T, Jarged S (2016) Virtualization techniques in cloud computing. *Imper J Interdiscip Res (IJIR)* 2:1476–1479
5. Subbaiah KV, Kiran Kumar P (2014) Dynamic resource allocation using virtual machines for cloud computing environment. *Int J Adv Res Softw Eng* 4:978–990
6. Frtala T, Zokova K (2014) Virualization: an answer to secure development of online experiments. The International Federation of Automatic Control, pp 9738–9743
7. Sharma AK, Soni P (2013) Mobile cloud computing. *Int J Innov Eng Technol (IJIET)*, 24–27
8. Kaushik N, Kumar J (2014) A literature survey on mobile cloud computing: open issues and future directions. *Int J Eng Comput Sci* 3:6165–6171
9. Oberheide J, Veeraraghavan K, Cooke E, Flinn J, Johanian F (2011) Virtualized in-cloud security service for mobile devices. Electrical Engineering and Computer Science Department, pp 1–5
10. Sunitha Rekha G (2018) A study on virtualization and virtual machines. *Int J Eng Sci Invention (IJESI)*, 51–55
11. Ghorpade Y, Bennur T, Acharya HS, Kamatchi R (2015) Server virtualization implementation: an experimental study for cost effective and green computing approach. *Int J Comput Sci Trends Technol (IJCST)*, 109–123

# Implementing Slowloris DoS Using Docker



Ishaan Sharma, Manohit, and Abhinav Bhandari

**Abstract** In this article, we are going to introduce the implementation of Slowloris attack on Apache web server running inside docker container. We will show how a server running inside docker container can be exploited. As we know, a server is a computer or a computer program that manages access to centralized resource or a service in a network and therefore is the core of the internet. One of the attack by which a server can be exploited is denial of service (DoS) Attack. The main aim of DoS attack is to shutdown a service or a network making it inaccessible to the intended users. Slowloris is a tool which is being used for DoS attack. Slowloris is a tool which lets single machine to take down web server with minimal bandwidth. Detailed implementation of this attack will be illustrated in this paper.

**Keywords** Slowloris · Docker · Denial of service attack · Ethical hacking · Apache server

## 1 Introduction

As the name suggests, the denial of service (DoS) attack is a kind of attack done to disrupt the services provided a machine and thus shutting it down. It is done by either flooding the network or sending the data that crashes the network or machine depending upon the situation. Flooding attack occurs when the system or network receives too much traffic than it can handle and eventually it slows down and crashes, which includes buffer overflow attack, SYN flood. The second type of attack is simply

---

I. Sharma (✉) · Manohit  
Bachelor of Technology, Punjabi University, Patiala, India  
e-mail: [ishaansharma.aries@gmail.com](mailto:ishaansharma.aries@gmail.com)

Manohit  
e-mail: [Kumarmanohit032@gmail.com](mailto:Kumarmanohit032@gmail.com)

A. Bhandari  
Punjabi University, Patiala, India  
e-mail: [Bhandarinitj@gmail.com](mailto:Bhandarinitj@gmail.com)

exploiting the vulnerabilities of the system and crashing it. The risk elevates when the attacker uses DDoS attack that is Distributed DoS attack in which multiple attackers run a DoS attack on a single system. DoS attacks are mostly used not only to shutdown a service but also to camouflage other cyber attacks. The most prominent target of DoS attacks is finance and gaming industries because of the services provided. For example, if the hackers want ransom from their victims, they need to disrupt that service that is always available to the users. Hackers always choose the target that gives them maximum profit and least resistance. Slowloris, a DoS tool first designed by Robert Hansen in 2009 which allow us to bring down a bigger machine by a single attacker machine by opening multiple socket connections than it can handle. To understand the slowloris, let us first understand an analogy about the working of HTTP/GET request. If we want to surf the internet, we first need to setup connection with the server, it is done by the client machine by sending SYN packets to the server, asking to setup a connection, then the server replies with ACK packets to establish the connection. Once the connection is established, the client sends a GET request to fetch a file and the server returns the file and closes the connection. That is how a normal request over the internet works. A server has a limit to many connections can it handle at a time, if more clients try to access it than it can handle the server goes down and starts dropping others connections. Denial of service attack always tries to reach this limit to force server to drop legitimate connection requests. That is what Slowloris DoS does but in a different way, rather than sending malicious information [1, 2], it sends partial/incomplete HTTP traffic that server does not close the connection after completing the request since the request never gets completed, and sending live headers after every few seconds so that the server does not timeout the connection, e.g. by saying the server that “hey, I am still here I am just really slow” and therefore reaching the maximum socket limit of the server, and hence, the server starts to drop all other legitimate request. The data sent over slowloris is completely legitimate therefore making it really hard to detect over the network [3, 4].

Docker is emerging as a forefront innovation for deploying software applications and operating system virtualization, first released in 2013 by Docker. Inc for most of the part by used by cloud companies. Most of the applications for testing are deployed on docker no matter what kind of program or software it is. Docker just works by pulling the image of a software with all dependencies required to run the software and then running it as container under docker platform. We in this article are trying to pull an image of an apache server (`httpd:2.x`) inside the docker and running a denial of service attack through slowloris.

University of Albany was assaulted with about 17 DoS attacks under the period of Feb 5–Mar 1, 2019, downing the services for around 5 min, not affecting the data but disrupting the services.

In early Feb, the website of National Union of Journalists, Philippines was also hit. The site was disabled for several hours by a series of powerful attacks, peaking the traffic at 468 Gb/s.

Also, in mid-March, Facebook encountered serious problems with its service when people were not able to login to their Facebook or Instagram accounts. Many observers considered the attack to be DDoS attack but not yet confirmed [5, 6].

## 2 Related Work

Within the course of time, many researchers have come forward to devise the strategic techniques of implementing or improvising denial of service attacks and also prevention techniques from these attacks. Some of them are listed below:

- Proposed By Gianluca Papaleo: Internet is flooded with DoS attacks including detection strategies. In the given reference paper the author analyses slow DoS attack on web applications and hence propose a taxonomy to categorize each attack. The proposal of author's work is to classify slow DoS attacks for better understanding of its action strategy and helping engineers to develop further better strategies for security. DoS attack has been evolved into attack which allows flooding and vulnerability attack simultaneously, making difficult for the network administrators to detect the malicious traffic; moreover, the attack can be launched through even a mobile phone. In this paper, the author analyses slow DoS attack to the web applications which use a low bandwidth rate to hit a web service by providing a proper taxonomy [7, 8].
- Proposed by Enrico Cambiaso: Denial of service attacks have evolved and have become a great threat to the organizations and governments all over the world. Earlier DoS attacks included high-bandwidth flood-based attacks. Subsequently, distributed DoS attacks evolved amplifying the consequences of the DoS attacks, thus bypassing the simple counter-measures. Slowloris is particularly dangerous; it can bring down a well-equipped server using small attacker's bandwidth, hence making them be able to run on low-performance hosts such as routers, game consoles and mobile phones.
- Proposed by Maurizio Aiello: Internet has become the hub for cybercrime and cyberwarfare. In this, the author presents the threat called SlowComm by showing how it can successfully lead a DoS on a targeted system using a small amount of attack bandwidth. Since this attack is not bounded to a specific protocol, it can be represented as a protocol independent attack, providing it the ability it has to effect different layers of internet [9].
- Proposed by C. L. Chowdhary: A reliable secure and platform-free remote controller, with ability of monitoring can overcome many problems. In this paper, a new design of network-based remote controlling and monitoring is proposed, which is platform free and more secure in comparison to other existing methods [10].

### 3 Implementation

Here, we are implementing Slowloris on an Apache 2.0 server running inside a docker container, it is done via Kali Linux terminal. To implement Slowloris, we need to download [11] and run docker in the linux shell. Since docker allows us to pull images from its directory of any software, so we will pull apache 2.0 software, but first we need to create or download web template and save it in any folder in local drive. Now, we need to pull the apache server from docker hub and map it to the folder we created for the web template by the following command,

```
$ docker run -d --name apc -p 8080:80 -v $(pwd)/ishwp:/usr/local/apache2/htdocs httpd:2
```

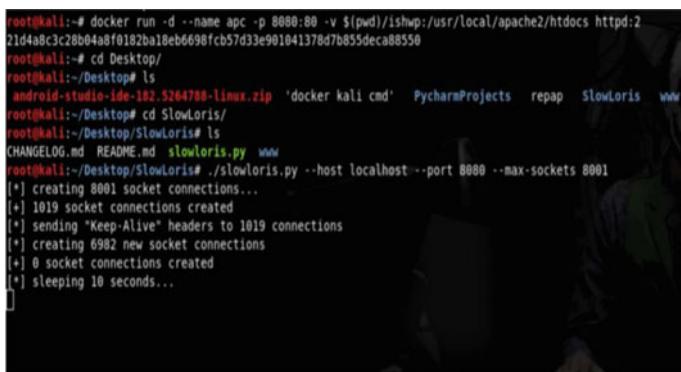
This command will pull and run the apache 2 container at port 8080 and will the folder in which the webpage resides (ishwp in this case) and hence the webpage will run at localhost:8080 until the container runs. Now, proceed to your Slowloris script and execute that code in at the localhost and at same port we are running the website, and a few moments later we can see the website crashes.

```
./slowloris.py --host localhost --port 8080 --max-sockets 8001
```

What this command does is that it runs the slowloris python script in the shell at localhost:8080 with socket creation of 8001 (as many as possible).

As we can see in Fig. 1 that first a running container is being created with a test website inside in it and then we scroll to our Slowloris script folder and implement it on our test webpage.

Figure 2 depicts the test website we are running the Slowloris attack on.



The terminal window shows the following sequence of commands and their output:

```

root@Kali:~# docker run -d --name apc -p 8080:80 -v $(pwd)/ishwp:/usr/local/apache2/htdocs httpd:2
21d4a8c3c2bb04a8f0182ba18eb6698fc57d33e901041378d7b055deca88550
root@Kali:~# cd Desktop/
root@Kali:~/Desktop# ls
android-studio-ide-182.5264788-linux.zip 'docker kali cmd' PycharmProjects repap SlowLoris www
root@Kali:~/Desktop# cd SlowLoris/
root@Kali:~/Desktop/SlowLoris# ls
CHANGELOG.md README.md slowloris.py www
root@Kali:~/Desktop/SlowLoris# ./slowloris.py --host localhost --port 8080 --max-sockets 8001
[*] creating 8001 socket connections...
[+] 1019 socket connections created
[*] sending "Keep-Alive" headers to 1019 connections
[*] creating 6982 new socket connections
[+] 0 socket connections created
[*] sleeping 10 seconds...

```

**Fig. 1** Terminal view of implementation

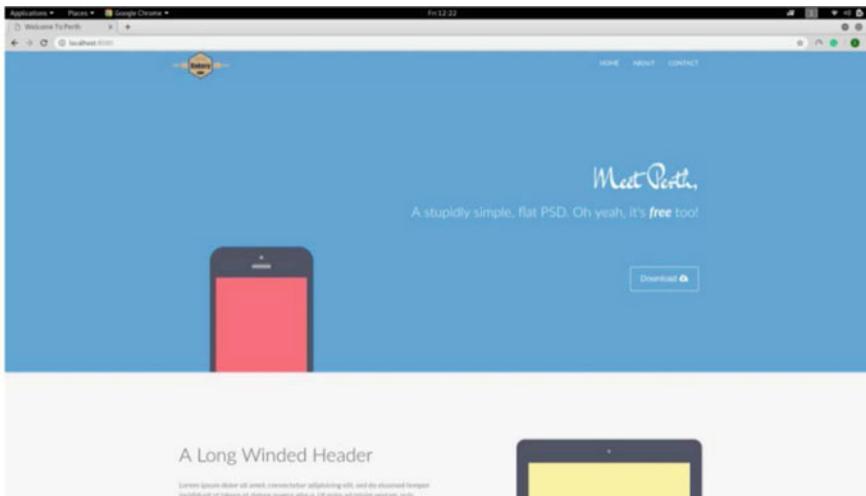


Fig. 2 Test website [12]

## 4 Topology

Figure 3 depicts the topology of the network, all working under the same machine. As we can see, we send partial HTTP/GET requests to the Apache Server container from our terminal to access the webpage (shown in red colour), these are the requests aren't actually fulfilled by the server, all running inside the docker, with containers (Nginx in this case). Then a legitimate request is sent at the same port on which the

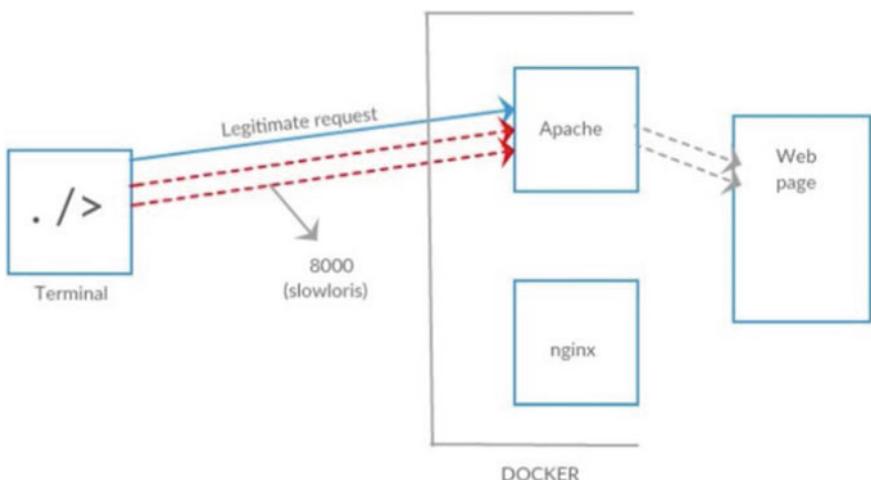


Fig. 3 Topology of network

webpage is running i.e. 8080 (shown in blue colour), which never gets completed (shown in grey colour) as the request queue of the apache server is full. The Slowloris sends 8000 partial HTTP requests which are never fulfilled by the attacker.

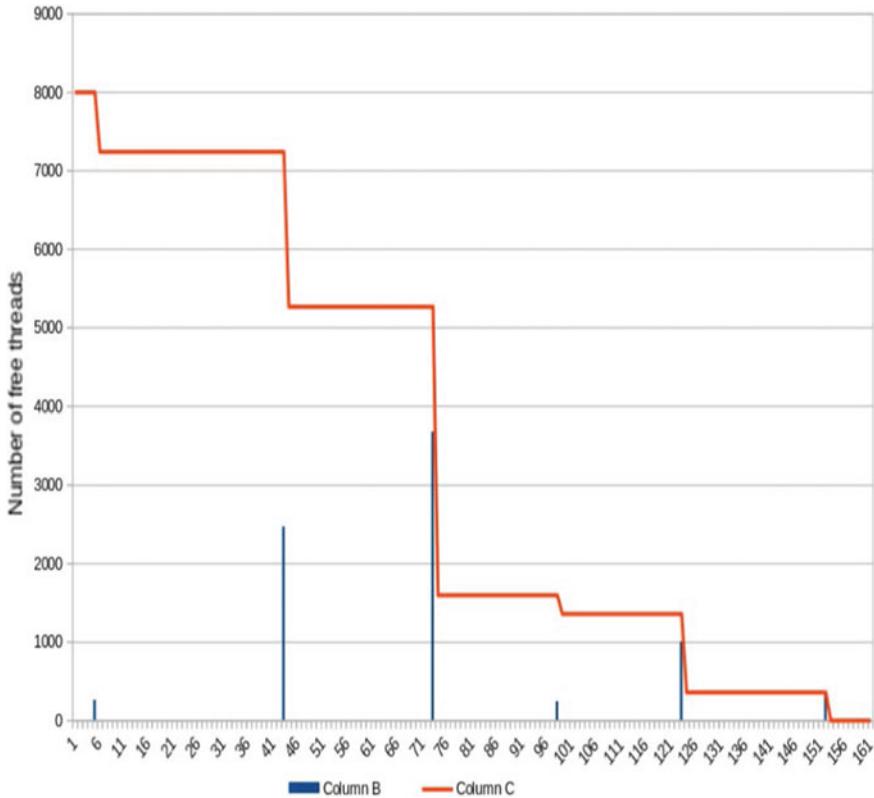
## 5 Results

What Slowloris actually does is that it starts creating multiple socket connections on the same port. We do not send complete HTTP requests because the server will shut down the connection after completing the request we send partial HTTP/GET requests which are extremely slow, thereby not letting complete connection to be established to the server. But here, we get a central issue, if we keep the server waiting for too long with extremely slow requests, the server will automatically shut down the connection showing “Connection timed out”. To overcome this issue, we keep sending live headers to the server every few seconds therefore notifying it to not to close the connections and therefore establishing multiple connections than the server can handle, ultimately slowing down and finally crashing.

But Slowloris has one major limitation that it works only for Apache 1.x,2.x server and other servers like Lighttpd and Nginx are not prone to the Slowloris attack, because it is the way how Slowloris works. As discussed above, the Slowloris will flood the server with connections and hence the queue will keep getting larger and eventually bring the server to halt using up all its memory, but in server like Nginx, a single thread can handle multiple concurrent connections unlike Apache where a single thread handles only one connection. Nginx uses an event-driven process for handling requests. Nginx and Lighttpd servers ignore requests with no ongoing processes and let them run in the background and handling requests which has something going on, e.g. accessing database at backend etc, thus making it quite invulnerable to the Slowloris attack.

As Fig. 4 depicts, the graph shows how the server goes down with respect to time. Column B represents the number of packets we send with partial HTTP/GET requests over a period of time of around 160 s before the server finally crashes. The column C represents how many threads are left in the server to handle the connections. We can see after nearly about a period of 155 s the server runs out of threads to handle the requests and finally crashes. The threads start from 8000 in the start and end in 0 at about estimated 155 s, so we can say that Slowloris takes about 2 and half minutes to bring the apache server running inside a docker container.

As we can see in the above image, TCP retransmission occurs from the docker IP which acts as a router (172.17.0.1) to the destination IP (172.17.0.2). The TCP retransmission occurs as we cannot complete the requests sent by the client. TCP transmission can happen only when the transmitting end does not receive TCP ACK from the receiving end, and the system is still trying to synchronize the packets by sending SYN requests, then it starts to continue this in the loop and finally there is transmission of RST packets as shown in adjacent figure in red colour. When one side sends RST packets, the socket is closed immediately and receiving side also

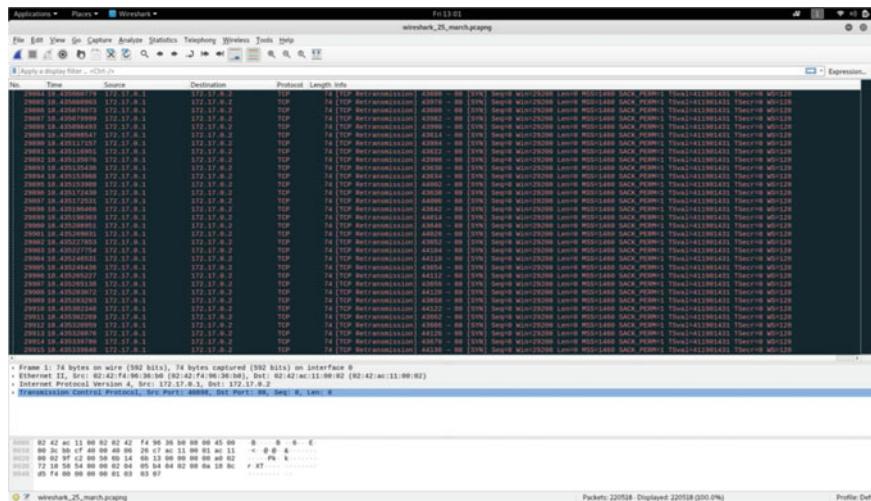


**Fig. 4** Graph depicting falling of server threads with respect to time

closes socket after receiving a valid RST request and hence the service provided by the server comes to halt (Fig. 5).

## 6 Conclusion

Now it is clear that a server running inside a docker can be hacked by DoS attack, thereby bypassing the security of docker as well. The image of Apache server running inside the docker is prone to DoS attack due to the flaws in the docker system as well, the slowloris by creating multiple half httpd connections makes all the server threads busy at the same time and thereby not allowing other legitimate users to connect to it. As research has demonstrated that Slowloris has emerged as one of the finest tools for DoS attack, due to its effectiveness, it can bring down the servers like Apache down the server by creating the multiple socket connections at the same port, which holds approximately 50% of the market share. It is clear from the above



**Fig. 5** Wireshark view

results that how Slowloris brings to nginx and lighttpd servers because they use worker threads instead of maximum number of connections, so there is no need of maximum connections for nginx and lighttpd.

First developed in 2009, its utility has ever grown since. Researchers are still working to improve the scale of this attack over the period of time. The other type of DoS attacks includes Teardrop/IP Fragmentation attacks, UDP flooding, SYN flood and Ping of Death attack [8].

Security Engineers are putting a great amount of effort to bring down such attacks, developing high end algorithms to counter these kinds of attacks but attackers always stay one step ahead to the security experts to fulfil their jobs, finding vulnerabilities and hence designing their exploits over the period of time.

## References

- White GB, White G (1996) Computer system and network security
- Corner D (1998) Internetworking with TCP/IP
- Kurose Ross Pearson Publications (2000) Computer networking: a top down approach
- Thakkar D (2017) Preventing digital extortion
- Recent attack news, <https://securelist.com/ddos-report-q1-2019/90792/>
- Top network attacks, <https://outpost24.com/blog/top-10-of-the-world-biggest-cyberattacks>
- Taxonomy of Slow DoS Attacks to Web Applications [6]. [https://link.springer.com/chapter/10.1007/978-3-642-34135-9\\_20](https://link.springer.com/chapter/10.1007/978-3-642-34135-9_20)
- Dafydd Stuttard and Marcus Pinto (2007) The web application-hacker's handbook
- SlowComm—design, development and performance evaluation of a new slow dos attack. <https://www.sciencedirect.com/science/article/pii/S2214212616300680>
- Chowdhary CL, Design and implementation of secure, platform-free, and network-based remote controlling and monitoring system. <https://ieeexplore.ieee.org/document/6208342>

11. Branon Dorsey Slowloris, <https://github.com/brannondorsey/SlowLoris>
12. Webpage design, <https://colorlib.com/wp/free-html-website-templates/>

# Sentiment Analysis of Pulwama Attack Using Twitter Data



Ranu Lal Chouhan

**Abstract** Microblogging sites like Twitter have become significant sources of real-time information during a disaster. Millions of tweets are posted on Microblogging sites like Twitter have become significant sources of real-time information during a disaster. Millions of tweets are posted during disasters. This algorithm is applied to twitter tweets to extract the sentiments of the public on such types of manmade disasters. In order to use microblogging sites effectively during disaster events, it is needed to summarize the large amounts of real-time non-situation information posted on twitter. In this study, non-situational tweets were analyzed which were posted during the recent disaster event of the Pulmawa attack. The proposed methodology is to develop a Gradient Boosting classifier using machine learning techniques to achieve better performance compared to support vector machine and random forest classifier to categorize various types of non-situation tweets collected during disaster into a set of different classes. Well-known sentiments were used for mining that exhibits eight basic emotions, that is, joy, Trust, fear, surprise, sadness, disgust, anger, and anticipation.

**Keywords** Sentimental analysis · Twitter · Natural language processing · Machine learning

## 1 Introduction

Nowadays, popular research area that has emerged is the sentiments analysis. Sentiment analysis is used to identify people's opinions, emotions, and feelings about a product, service or event. It is computational research expressed in the text of opinions, emotions, attitudes, views, sentiments, etc. This text can be available in various formats such as Reviews, Blogs, News or comments. The fast development of Microblogging sites converges with those of the web-based social media, such as forum conversations, blogs, reviews, and Twitter, etc. because we have an

---

R. L. Chouhan (✉)  
Government Engineering College, Bikaner, Bikaner, India  
e-mail: [chouhan.ranu@gmail.com](mailto:chouhan.ranu@gmail.com)

enormous amount of judgmental information collected in digital form. These sites show a big quantity of useful information of review, opinions, sentiments, emotions, and attitudes toward entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1, 2]. Microblogging sites like Twitter are becoming one of the popular platforms of expressing sentimental on various issues like implementation of the government scheme and review consumer products etc. In the US presidential campaign of Barack Obama has established Twitter, Facebook, MySpace, and other social media as integral parts of the political campaign toolbox [3]. The Indian PM Modi is very active on twitter, and it has been using as a most impressive tool for connecting people and publicizing government schemes. Twitter sentiment analysis is one of the experimental tasks often used for various reasons, which aims to automatically define the polarity (i.e., positive, neutral or negative) of the tweet using natural language processing methods [4]. The sentiment analysis aims to understand different views of the public, on a particular issue into classes such as positive or negative.

This study represents the sentiments of the public on the most terrible terrorist attack for India that recently occurred in Pulmawa. Different kinds of non-situational tweets need to be separated for such study, a task that requires being automated elevated at high speed at which tweets are published. In this paper, we propose to a classifier based on Stochastic Gradient Boosting to separate distinct categories of non-situational tweets automatically. The classifier depends on a suggested set of tweets' low-level lexical and syntactic characteristics and works considerably better than a Random Forest classifier and SVM.

To our observation, this research is the first to characterize non-situational tweets that are posted on Twitter during the Pulmawa attack that showing anger on this security threat and sympathy or organizing charities.

The expected outcome from this research is improved performance of sentimental analysis of such type of disaster event.

## Sentiment Classification Levels

Sentiment classification is divided into three levels:

1. **Document-level Sentimental Analysis:** Sentiment is separated from the entire study in this operation, and here, we are more interested in the overall of the text. A whole opinion is described in light of the opinion holder's overall feeling. It categorizes an opinioned document as expressing a positive or negative opinion. It operates best when a single person writes the document and communicates sentiment on a single entity. The entire document is considered as the fundamental unit of data, and the document is considered and contains views on a single entity.
2. **Sentence level Sentimental Analysis:** A single document may have a different opinion about the same entity, but it is assumed that each sentence has a single opinion.
3. **Aspect/Feature-based Sentimental Analysis:** Sometimes, it is not enough to tell whether a text has a "positive" or a "negative" feeling. People may express

their sentiments as positive, neutral, or negative polarity about the event occurred or product.

**Sentiment Analysis Techniques:** These are mainly divided into two categories Lexicon-based and Machine learning approaches. With Lexicon, clustering algorithms are used to finding the sentiment of the input text. This technique does not provide context meaning. To remove this issue, a corpus-based approach is being used. The corpus-based technique depends on co-occurrence patterns in large corpus. You will also require a large corpus to get better performance.

## 2 Related Work

Many researchers worked to find out the public's sentiments in various areas. Liu [5] depicts that determining review is helpful for products and services that have a larger number of reviews, but ranking must reflect the fair distribution of positive and negative reviews. Ranking all positive reviews at the top is not a good idea. Lexical and syntactic components are used [1] for a novel grouping summary scheme for disaster-specific situational data. Brynielsson et al. [6] present an approach for collecting a lot of crises related tweets and labeling relevant tweets using human annotators. The method was used to comment on significant quantities of tweets and the resulting information was used to build classifiers prepared to acknowledge the emotional classes as an outcome. It suggests that an SVM classifier is superior to an NB classifier. Vieweg et al. [7] present a model for enhancing situational awareness in crisis situations through information provided by those affected. Data generated during crises are used to develop a working framework. Such frameworks can be used by people from the general population and crisis respondents in their tasks to increase situational awareness in times of crisis. Nguyen et al. [8] present a model TSUM4act, for recovering enlightening tweets in the midst of a sensible reaction calamity. This model uses machine learning methods, event extraction, and graphical model to handle vast quantity tweet's noise and diversity during a disaster event. Ganguly et al. [9] proposed a model using the Hadoop cluster and Naive Bayes method to process and evaluate an enormous quantity of real-time data. The document focuses on the velocity of the assessment of sentiment rather than the accuracy of the assessment. It processes the input data by removing stop words, converting unstructured data into structured information and replacing sentiments with their respective phrases. Varga et al. [10] use bag-of-words classifiers to differentiate between tweets situational data and non-situational tweets. The classifier is trained on tweets of some previous occurrences in all real-time disaster events. Grover et al. [11] present a model using machine learning in which the framework will be trained to capture and react to influenza-like disease events for which preventive steps can be taken to get rid of the epidemic as soon as possible. Using the Delphi technique, the results provided by the Swine Epidemic Hint Algorithm are evaluated toward the ground Truth. Ain et al. [12] neural networks can result in more accurate outcomes

than Naive Bayes and maximum entropy, but it requires large datasets which is very costly to train. This works better for image processing but it requires expensive GPUs. Goel et al. [13] share the study that how disaster situations can be better managed by the participation of affected people during disaster. The information shared on the social media platform plays a very vital role in disaster management. Karami et al. [14] propose a framework that utilizes text mining techniques for producing a better situational awareness during any event of a disaster.

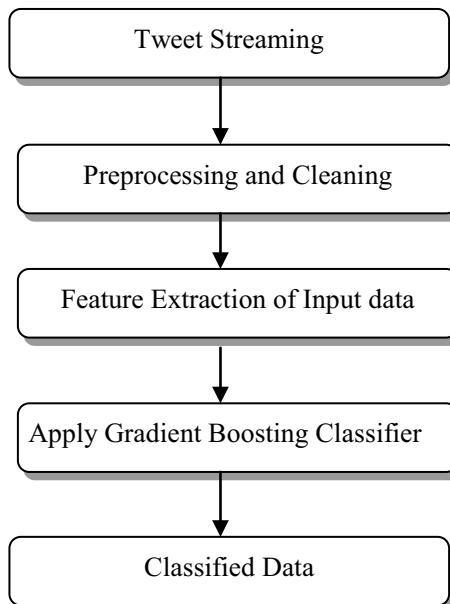
## **2.1 Research Methodology**

The model presents the steps required to classify the non-situational tweets during Pulwama Attack. Our purpose is to study only non-situational tweets about “2019 Pulwama Terrorist Attack,” so we have considered 4000 non-situational tweets only.

## **2.2 Data Collection and Preprocessing**

All the tweets and retweets have been collected about Pulwama Attack through the streaming of Twitter API as shown in Fig. 1 Dataset has been prepared by Twitter Application. Python libraries like tweepy, nltk, Sci-kit, etc. have been installed for

**Fig. 1** Basic framework of sentiment analysis



registering Twitter Application to get credentials like consumer key and consumer secret as well as token key and token secret. Twitter streaming code has been created to download tweets based on search keywords. We randomly sample 15,000 English tweets and preprocessing these tweets. Our dataset may have any unwanted data, noise or non-opinion words that do not provide any useful information. So we have removed following punctuation marks, stop words such as about, almost, etc., twitter specific terms such as URLs, #, @, etc., special symbols and emoticons symbols. We have normalized whitespace, converted all uppercase to the lowercase letter also.

## 2.3 Feature Extraction

**Separating situational and Non-situational Tweets:** Different types of tweets are collected using twitter API, these include some tweets for situational awareness and other can be called non-situational tweets. These tweets have been separated by classifier [1] as well as human volunteers having good knowledge of English and twitter.

These Tweets can be classified into following two broad categories:

1. **Situational tweets:** These tweets are related to real-time situation updates. These tweets provide useful information about factual of the affected area during a disaster. What is happening in the affected area? This information helps to affected people during a disaster.
2. **Non-situational tweets:** These tweets are not related to any real-time situation updates or immediately help the rescue operation. These include General information like suggesting, criticizing, appraising, charities, expressing anger or consolation, etc.

### Classification of Non-situational tweets [1, 15]

**Sentiment tweets:** Criticizing government agencies, organizations, individuals and sympathizing with the disaster victims, and praising for victims.

**Opinion tweets:** opinion or suggestion given to government agencies/organizations/individuals for improving the rescue operation.

**Political tweets:** Tweets about condemn of some political parties and criticizing some political leaders/parties.

**Event analysis:** Post-analysis of how this disaster happened and how could this disaster be prevented and whose mistake or negligence was there in this disaster. Finding or trace the investigation of this event.

**Organizing charity:** Tweets related to organizing the charity to help the victims.

**Emotion related tweets:** Expressing only personal feelings without any above-given tweets like anger, sadness, anxiety, or providing any emotional support, encouraging and expressing consolation for victims.

**Communal tweets:** Any tweet which attacks any religion, a community for occurring of disaster.

**Table 1** Example of various non-situational tweets posted during “2019 Pulwama Terrorist Attack”

Feature Counting of words	Example of tweet text
Sentiment	@narendramodi We salute you, brave sons of Mother India. You lived for the nation and served the country with unparalleled valour. We stand in solidarity with the bereaved families
Opinion	@paavanichalla Another surgical strikes please.....modi ji thus great ho
Political	@abhijitmajumder Remember. Anyone on your timeline saying CRPF jawans died because of “political failure” in Kashmir is trying to divert attention from the foremost threat to civilised world today: Islamist terror
Event Analysis	@roshankr How such huge explosive came
Charity	@republic Amitabh Bachchan to donate Rs 5 lakh each to families of soldiers killed in Pulwama attacks #StandWithForces
Emotion	@nikhengr Why v haven’t shut down Pakistan’s embassy yet????
Communal	@HamzaAliAmjad I am Muslim Buddhist kill me in Burma. Christians kill me in Afghanistan Hindu in Kashmir Jewish in Palestine. Still I am terrorist! #Pulwama
Offensive	@aarifshaah Indian hotels allowed ‘dogs’ but not ‘Kashmiris’ in their hotels after Pulwama attack #pulwama

**Offensive tweets:** Any tweets like insulting, disrespectful, abusive or to make fun of victims.

**Off-topic tweets:** Any tweets irrelevant to this particular disaster (here-“2019 Pulwama Terrorist Attack”) (Table 1).

## 2.4 Classification

We have developed a Gradient Boosting classifier to differentiate different types of non-situational tweets. Lexical-based classifier does tokenization into lexemes. The lexemes are matched with manually created words of dictionaries. We have made the NLTK based function for feature extraction. This dictionary consists of all class of tweets that are available in our dataset. We have created and annotated a corpus of 4000 non-situational tweets by the help of five volunteers, and it will be distributed among eight different classes (Table 2).

**Train test split:** We have only one dataset of “2019 Pulwama Terrorist Attack,” so we have to randomize the training classifier on different datasets. We have split

**Table 2** Percentage of various non-situational tweets

Sentiment	Opinion	Political	Event analysis	Charities	Emotion	Communal	Offensive
42.26	16.52	5.26	7.26	12.43	9.25	3.28	1.85

**Table 3** The Performance result of various classifiers

Classifier	Accuracy
Support Vector Machine	79.92
Random Forest	81.26
Gradient boosting classifier	82.45

our corpus into training and testing sets and used a 75% training set to train the model and 25% test data to test the model. Our 4000 non-situational tweets have been classified among eight classes. We have used 5-fold cross-validation to tune the hyperparameter for obtaining better accuracy of the model. The Test set was used to optimize our proposed approach. The target of designing a “gradient boosting algorithm” is to keep constructing a tree with low bias (Table 3).

### 3 Results and Discussion

The tweets related to the “2019 Pulwama Terrorist Attack” have been analyzed to understand the sentiments of the public with respect to random forest, SVM and our Gradient Boosting classifier. The experiment found that random forest classifier performs better only in sentiments and opinion class but our proposed model works better for all class of tweets. These public sentiments may be useful for auctioning taken by the government. In disaster events, offensive tweets that may harm a piece of society may be extracted in real-time situation and action may be taken against him.

### References

1. Sen A, Rudra K, Ghosh S (2015) Extracting situational awareness from microblogs during disaster events. In: 2015 7th international conference on communication systems and networks (COMSNETS). IEEE
2. Rudra K et al Extracting and summarizing situational information from the twitter social media during disasters. ACM Trans Web (TWEB) 12(3):17
3. Tumasjan A et al (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Fourth international AAAI conference on weblogs and social media
4. Wehrmann J, Becker WE, Barros RC (2018) A multi-task neural network for multilingual sentiment classification and language detection on Twitter. In: Proceedings of the 33rd annual ACM symposium on applied computing. ACM
5. Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167
6. Brynielsson J, Johansson F, Westling A (2013) Learning to classify emotional content in crisis-related tweets. In: 2013 IEEE international conference on intelligence and security informatics. IEEE
7. Vieweg S et al (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM

8. Nguyen M-T, Kitamoto A, Nguyen T-T (2015) Tsum4act: a framework for retrieving and summarizing actionable tweets during a disaster for reaction. In: Pacific-Asia conference on knowledge Discovery and data mining. Springer, Cham
9. Ganguly M, Roy S (2018) A social network analysis of opinions on GST in India within Twitter. In: Proceedings of the workshop program of the 19th international conference on distributed computing and networking. ACM
10. Varga I et al (2013) Aid is out there: looking for help from tweets during a large scale disaster. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers)
11. Grover S, Aujla GS (2015) Twitter data based prediction model for influenza epidemic. In: 2015 2nd international conference on computing for sustainable global development (INDIACoM). IEEE
12. Ain QT et al (2017) Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl* 8(6):424
13. Goel A, Chakraborty M, Biswas SK (2019) The role of social media in crisis situation management: a survey. In: Emerging technologies in data mining and information security. Springer, Singapore, pp 439–448
14. Karami A et al (2019) Twitter speaks: a case of national disaster situational awareness. *J Inf Sci*, 0165551519828620
15. Qu Y et al (2011) Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. ACM

# A Survey on Architecture and Protocols for Wireless Sensor Networks



Anita Chandel, Vikram Singh Chouhan, and Dhawal Vyas

**Abstract** Wireless sensor networks (WSN) are employed in many application areas such as environmental monitoring and battlefield strategy planning. Mainly, wireless sensor networks(WSN) are application specific. However, protocol-specific requirements of WSN are same as traditional networks, but the solutions are different. Many existing WSN implementations do not address security requirements, routing algorithms, MAC etc. That is, a drawback of WSN network protocol. WNS sensor nodes are constrained by memory, bandwidth, and power requirements, which makes it difficult to deploy complex algorithms, perform bulky calculations, and store a large data set in any sensor node. In this paper, we will describe the architecture and protocols, generally used to provide a variety of services in WSN. The introduction is started with architecture and principles of wireless sensor networks. The next section describes protocol stack for WSNs. After that, there is a brief survey of various protocols categorize on the based on layers and services.

**Keywords** Wireless Sensor Networks · WSNs · Architecture · Protocols · Layers · Applications · Network Models · Resource Constraints · Self-Organization

## 1 Introduction

A wireless sensor network (WSN) is a wireless network, which consists of distributed autonomous nodes (sensors). Nodes are capable of capturing information, processing, and communicating to each other, and routing data back to the base station. Base

---

A. Chandel (✉) · V. S. Chouhan

Information Technology Department, Engineering College, Bikaner, Bikaner, India  
e-mail: [anita2506@gmail.com](mailto:anita2506@gmail.com)

V. S. Chouhan

e-mail: [vikksecb@gmail.com](mailto:vikksecb@gmail.com)

D. Vyas

Information Technology Department, Engineering College, Bharatput, Bharatput, India  
e-mail: [d2vyas@gmail.com](mailto:d2vyas@gmail.com)

station issues commands to manage sensors, collects sensors data, and processes it. WSN provides a simple and economic approach for the deployment of distributed sensing nodes and controlling them. WSN plays important role in a wide variety of area ranging from sensing physical and environmental conditions, measuring medical parameters for health applications, to security-related application as in hostile environment of a battlefield, building security, fire fighting, etc. A WSN consists of many nodes which are either fixed location-based sensors or movable sensors in the monitoring environment.

Due to their size and deployment scenario, sensor nodes have a number of limitations. A sensor/node contains two kind of memory,

- i. **Flash** stores application, configuration parameters
- ii. **RAM** stores sensor data, intermediate results of computations, keys used for communication security.

Usually, there is not enough space to run complicated security algorithms.

Further, WSNs are dynamic by nature. Sensors die and new sensor might be installed to the sensor network. Due to this, the topology of WSN keeps changing. The positions of sensors are not predetermined, and hence, WSN needs to have self-organizing capabilities.

The nodes in WSN have resource constraints due to less memory, energy, processing power, communication capabilities, and frequently changing topology. To better understand why traditional network protocols are not suitable for these types of sensor network applications, the unique features of sensor networks are categorized in further sections.

## **1.1 Applications of WSN**

There are two types of nodes:

1. **Source:** Senses the data.
2. **Sink:** Where data should be delivered.

The interaction between source and sink is application dependent. There are a variety of applications in which WSN can be categorized.

1. **Periodic Tasks:** In this type of applications, the sensors report periodic values after some time stamp for a given time duration. Example is temperature reading in remote areas.
2. **Event Based:** Sensor nodes sense the data and forward to the sink when any particular event occurs. All the events should be defined in prior.
3. **Tracking:** When the object is mobile, then to sense the location, speed or duration; the node will track the object to generate the desired results.
4. **Approximation Based:** When WSN is not able to produce the exact result or information, rather have some samples of data, then the result will be produced by approximating the event results.

The applications can also further categorized on the basis of deployment criteria, maintenance, energy supply, processor speed etc.

## 1.2 Challenges to WSNs

Due to the basic nature of WSN, this type of networks faces some challenges, which don't exist in traditional networks.

1. **Application Specific:** WSNs are application dependent and do not provide solutions like one fit for all types. The networks or even nodes are designed such a manner that support the application and produce favorable or desired results.
2. **Unattended or Hostile Environment:** WSN nodes are deployed in hostile environment which are unattended by human.
3. **Scale:** The size of a WSN is thousands or even hundreds of thousands of the sensor nodes. To provide a unique ID to a node is not feasible and hence, create a space for different scalable solutions.
4. **Energy Aware:** Sensor nodes contain small battery and replacing or recharging the battery is not possible for WSN. This issue makes the WSN operations more energy efficient.
5. **Self Configuration:** WSN is self-configured in the connected network, based on the traffic, remaining battery power, etc.
6. **Simple and Resource Aware:** The architecture of the sensor node is simple and resource constrained. Therefore, its operations, networking, communication, and computations must be kept simpler as compared to the traditional networks.
7. **Mobility:** The mobility causes by nodes moving around and change its locations with respect to the previous one. Due to mobility, the change will reflect in multi-hop routes and geographic topology. There are three types of mobility defined in WSN: Node mobility, Sink mobility, and Event mobility.
8. **Fault Tolerance:** Individual nodes are prone to unexpected failure with a much higher probability.

## 1.3 Network Models

A WSN generally consists of few base stations and hundreds or thousands of sensor nodes. Base stations have all the capabilities of battery powered, large memory space, powerful data processing, gateway to other networks, and direct interaction with human, as a traditional computer do. Whereas, Sensor nodes are quite constrained in all the aspects. Sensor node senses the data and collectively sends to the base stations. Being a better resource equipped, base station performs all the communication and computations to produce the required results.

WSN architecture supports two different types of setting related to network modeling:

1. Hierarchical versus Distributed model.
2. Homogeneous versus Heterogeneous model.

## 1.4 Basic Hardware Architecture

WSNs are completely dependent on the hardware architecture of its components. There are five main components of a WSN hardware:

1. **Controller:** It gathers required information from the sensor nodes, forms these information, and chooses which and when information ought to be sent or got. An assortment of handling undertakings can be performed by the controller are speaking to the exchange offs between adaptability, execution, energy proficiency, and expenses. The arrangement is the processors, normally known as micro-controller. Intel StrongARM [1], Texas instruments MSP430 [2], and Atmel ATmega [3] are few widely used micro-controllers.
2. **Memory:** Distinct types of memories are used in sensor node on the base of the data to be stored.  
RAM (Random Access Memory) is utilized to store moderate sensor readings, bundles from different sensors, etc. Program codes can be put away in ROM (Read-Only Memory), EEPROM (Electrically Erasable Programmable Read-Only Memory) or flash memory. Flash memory is also used for intermediate data storage.
3. **Communication Devices:** There are mainly two types of the communication devices:
  - (a) **Transmission Medium:** For the transmission, WSN generally opt between radio frequencies, optical communication, ultrasound, and other medium like magnetic inductance or Bluetooth in very specific scenario.
  - (b) **Transceivers:** For communication, both a transmitter and a receiver are required. It is convenient to combine both in a single sensor node, known as transceivers. Transceivers work in half duplex mode. Transceivers must account some characteristics as provide services to upper layers, power consumption and energy-efficiency, carrier frequency and multiple channels, state change times and energy, data rate, modulations, coding, transmission power control, noise figure, gain, power efficiency, receiver sensitivity, range, blocking performance, out of band emission, carrier sense, frequency stability, and voltage range. There are few examples of radio transceivers like RFM TR1000 family [4], hardware accelerators (Mica Motes) [5], chipcon CC1000, and CC2420 family [6], Infineon TDA 525x family [7], IEEE 802.15.4/Ember EM2420 RF transceiver [8], National Semiconductor LMX3162 [9], and conexant RDSSS9M [10].  
**Sensors and Actuators:** Sensors are further categorized in two parts:
    - i. **Passive:** These sensors sense or measure a quantity without manipulating the environment by active probing, hence known as passive. On the basis on direction, it can be sub-divided into omni-directional and narrow beam.
    - ii. **Active:** These types of sensor nodes actively probe the environment.

**Actuators:** Actuators are similar to sensor node but a bit simpler. Sensor node and actuators can be used interchangeably.

4. **Power Supply of Sensor Nodes:** There are two important aspects related to power supply. To start with, putting away energy and giving force in the necessary structure. Second, endeavoring to revive devoured energy. To store the energy, small batteries are used. For energy rummage, several approaches exist as photovoltaics, temperature gradients, vibrations, pressure variations, and flow of air/liquid.

## 1.5 Design Principles for WSNs

Accurate or efficient QoS support, energy efficiency, scalability are few prime design and optimization goals to achieve for a good WSN [11, 12]. Some basic principles are already emerged in WSNs when network protocols are designed, but others appended as per requirements.

1. **Distributed Organization:** Scalability and robustness optimization imposed to organize the network in a distributed fashion. Due to various disadvantages of centralized approach, the WSN nodes opt for cooperatively organizing the network using distributed algorithms and protocols. This principle is also known as Self-Organization.
2. **In-network Processing:** In self-organizing, the nodes accomplish their basic tasks like passing the packets or executing any application, but it must actively participate in decision-making to manage the network. For this purpose, an in-network processing design required. There are some existing techniques for in-network mentioned below:
  - (a) **Aggregation:** In aggregation, intermediate nodes aggregate the information in abbreviated form, derived only from the data it sense. This compact information further forward to the sink.
  - (b) **Distributed source coding and distributed compressions:** When the network is sufficiently dense, the information collected by adjacent nodes are quite similar. In such a case, correlation is used to reduce the overhead. Degree of correlation is also derived for some distinct nodes, known as temporal correlation.
  - (c) **Distributed and Collaborative Signal Processing:** In-network processing is adequate for only trivial operations like averaging or find min/max. Computing complex operations still can be more energy efficient. In this case, the concept of distributed computations arises, which is similar to algorithm design for parallel computers.
  - (d) **Mobile Code/Agent-based Networking:** To execute the program in network, programming paradigms or computational models are required. Such model is the concept of mobile code or agent-based networking. The

program code is small and compact enough to be sent from one node to other and executed locally.

3. **Adaptive fidelity and Accuracy:** The idea of the devotion can and ought to be reached out from single node to whole system. For Approximation, when more sensors are participating, the function samples more points and produce a better result.
4. **Data Centric:** In WSN, it is of no concern that data come from which node or which node is providing the sense data. Hence, the data are the center of attention, not the identity. So this approach is known as data-centric networking. There are three implementation options for data-centric networking: Overlay networking and Distributed hash tables, Publish/subscribe, and databases.
5. **Obtain Location Information:** The location of an event or a node is a crucial information for many application. Various mechanisms are prevailed to determine the location and mitigate to design and operation of communication protocols.
6. **Obtain Activity Pattern:** When an event has occurred, it very well may be seen by an enormous number of sensors. Hence, the design of a protocol should handle such burst of traffic and able to perform the specific action.
7. **Utilize Heterogeneity:** In WSN, heterogeneity can be observed in construction and evaluation. Due to this, asymmetry prominent for assigning the task to better-equipped node. This asymmetry leads to re-evaluation of tasks at the time passes. On advantageous side, better-equipped nodes have the assigned tasks of monitoring and perform complex operations on behalf of other sensor nodes.
8. **Component-based Protocol Stacks and Cross-Layer Optimization:** Due to application dependency, layered protocol stack is not feasible for WSNs. Based on the application, component-based stack and cross layer protocols are required. These are further elaborated in upcoming sections.

## 2 Structure of Operating System and Protocol Stack

In classical approach of communication protocol, layering architecture is used. Individual protocol and individual layer are stacked on the top of each other and provide and use services to upper and lower layer, respectively. Cross-layer information exchange is utilized to release the exacting imprisonments of the layered methodology. Departing from the layered architecture, component-based model commonly used in WSNs. Monolithic layers are separated into little and independent modules. These components satisfy only one well-defined function as per requirement. The main difference between traditional approach and cross-layer is the collaborations are not bound to prompt neighbors in a relationship, be that as it may, can be with subterranean insect different parts.

The component model is able to solve some structuring problems and fits naturally with an event-based approach to programming WSN. Wrapping of hardware, communication primitives, and in-organize handling functionalities helpfully structured and executed as segments.

### 3 Protocols in WSNs

In traditional wired networks, the system is very complex due to assembly of various components. These components have some set of functions which are defined collectively. At the core of these functions is a set of rules for information exchange. These rules are known as protocol. The systems are divided into chunks of well-defined functions. This function module is collaboratively known as layer in communication system. Each layer has assigned a set of similar tasks, distinct from other layers.

For the purpose of WSNs, it is not quite obvious that a strict layered architecture is sufficient or other approaches can be required. In WSNs, one single wellspring of data can be utilized to give points of interest to numerous different layers, which are not legitimately connected with the wellspring of data.

Such cross-layer information exchange is able to slack the strict captivity of layered architecture. Even in WSNs, traditional wired layered protocols can be used with the support of efficiency considerations. TCP over wireless is the best example of such systems. Gigantic layers are separated into little and independent parts. The component interaction is not completely bound with immediate neighboring layers, but it can interact with any other component. It solves some structuring problems for protocol stack and supports application-oriented approach of WSNs.

However, a WSN is a component-based cross-layer information exchange protocol, the study of protocol is still best suited in layered fashion. The components are sub-divided according to the layering characteristics.

#### 3.1 *Physical Layer*

Physical layers and transmission channels are responsible for shaping of the basic protocol stack. The physical layer is mainly concerned with modulation and demodulation of digital data, which are performed by transceivers. The prime concern is to find modulation schemes and transceiver architectures which are simple, low-cost, and robust.

However, there is no specified component-based protocol for physical layer but the studies are divided into few basic categories to implement for desired services. These categories are wireless-channel and communication fundamentals, and transceiver design considerations. Wireless-channel and communication fundamentals are mainly concerned with frequency allocation, modulation/demodulation, wave propagation, effects and noise, reflection, diffraction, scattering, Doppler fading, path loss, attenuation, interference, symbols and bit error, channel models, spread spectrum communication, packet transmission and synchronization, quality of wireless channels, and measures for improvements. The prime concern of transceiver design considerations is energy usage, choice of modulation scheme, dynamic modulation scaling, and antenna description.

### **3.2 Data Link Layer**

As per traditional network protocol, the link layer services are reliable delivery, flow control, error control, multiplexing of data stream, and link management. It uses services from physical layer and provides services to network layer like wired layer architecture. This layer is further divided into Medium-Access Control sublayer and Logical-Link Control sublayer. There exists a variety of protocols for link layer in protocol suite for WSN.

### **3.3 MAC Protocol**

MAC protocol [13] in WSN should be specialized enough to handle the issue of power conversion and data-centric routing. The major assignment of MAC convention is to direct the entrance of various hubs to a mutual medium and fulfill some application-arranged execution prerequisites. Delay, throughput, and fairness are the performance criteria in traditional network, whereas, the issue of energy conservation is prime concern in WSNs.

MAC protocol for WSNs narrow down the specific requirements and design considerations.

1. **Balance of requirement:** Additional requirement of energy conservation to design a MAC protocol is naive, and traditional approaches like ALOHA and CSMA have no provision to achieve this. Other important requirements for MAC protocols are scalability and robustness against frequent topological changes.
2. **Energy problem on the MAC sublayer:** Node's status can be categorized in one of the four stages: Communication, sensing, idle, or sleep. Communication is the costly one, and sleep is the cheapest one. Based on these rules, energy problem can be solved to design goals for MAC sublayer. Collision, overhearing, protocol overhead, and idle listening are main energy problem faced by a WSN. A rough classification of MAC protocol is fall into three categories: Fixed-assignment protocol, demand-assignment protocol, and random-access protocol.

### **3.4 Low Duty Cycle Protocol**

It tries to avoid the wastage of time and energy in the idle state and reduce communication activities. Several protocols use periodic wakeup scheme. By a small duty cycle, to avoid idle listening, the node can be in sleep stage. After a given time stamp, the node wakes up, senses, communicates, and again goes to sleep mode.

1. **Sparse topology and energy management (STEM):** It targets network, which wait and report of a certain event. The network has a monitor state, where node

- is idle and transfer state, where node senses and communicates, and eliminates idle listening.
2. **S-MAC:** Sensor MAC [14] adopts a periodic wakeup scheme, it alternately oscillates between fixed-length listen and sleep period, known as schedule. It has three phases: SYNCH phase, RTS phase, and CTS phase.

### **3.5 Contention-Based Protocol**

In this protocol, any node of the sink's neighbors may try for the risk of collision. These protocols contain a few systems to keep away from impact or to diminish their likelihood. CSMA and PAMAS [15] are the examples for this protocol.

### **3.6 Schedule-Based Protocol**

In this protocol, a node knows its slot of communication and makes sure to perform it only in that slots. In the remaining time, the node can simply switch off its receiver for safety issues. LEACH [16], SMACS [17], and TRAMA [18] are few examples of this protocol.

### **3.7 IEEE 802.15.4 MAC Protocol**

It covers the physical layer and the MAC layer of a low rate WPAN. Its targeted applications are home automation, home networking, home security, and so on. Slotted CSMA-CA protocol falls in this category of protocol.

### **3.8 Link Layer Protocol**

The logical link control (LLC) layer places on the top of MAC sublayer, which offers packet transmission and reception service. LLC gives administrations to the network layer and other higher layers.

The most significant errands of LLC layer are to make a solid correspondence, framing, error control, flow control, and link management. In framing, the data are fragmented and assembly at sender and receiver side, respectively. In error control, data transportation can be categorized in the terms of error-free, in-sequence, duplicate-free, loss-free, delay, and energy constraints. The most significant error control systems are forward error correction (FEC) and automatic repeat

request(ARQ). There exist a variety of protocols for FEC and ARQ as per requirements. In link management, discovery, setup, maintenance, and tear down of links to neighbors mechanisms are involved.

### ***3.9 Time Synchronization***

An evidently small error of time can lead to significantly biased estimations. This problem can be solve by keeping sensor clock tightly synchronized, known as time synchronization or by averaging the estimations error by combining the reading of various sensor nodes. The key example is MAC protocol based on TDMA or MAC protocol with coordinated wake up.

The protocols are categories into two classes: Sender/receiver synchronization protocols and Receiver/sender synchronization protocols. In first protocol, the time will be synchronized by sender's clock, whereas, in second one, it is by receiver's clock. Lightweight time synchronization protocol [19] (LTS) and timing-sync protocol for sensor network(TPSN) [20] are two examples of sender/receiver synchronization protocol, and reference broadcast synchronization(RBS) and hierarchy referencing time synchronization(HRTS) are of receiver/sender synchronization.

### ***3.10 Localization and Positioning***

It is useful for a node to find its location in physical world. Manual configuration of location of each node at deployment is not possible, and GPS is not suitable for WSN due to cost and deployment limitations. Physical position, absolute vs relative coordinates, computation-based location, accuracy and precision, scale, limitations and costs are few properties of localization and positioning procedure. Three possible approaches can be used to determine a node's position are as follows: proximity, trilateration and triangulation, and scene analysis. Positioning is distinct in the case of single hop and multihop.

### ***3.11 Network Layer***

In traditional wired layer architecture, the functionality of network layer to help in delivery of data in the form of packets from source host to destination host. Network layer helps to send data over different network by using logical and physical addressing. Routing is the prime concern of network layer by using switches and routers to route the packets to final destination.

The network layer in WSN must be structured with some thought of wired layered design with control effectiveness. WSN is data-centric network, rather than address-centric. WSNs are based on attribute-based addressing and location awareness. The data may be aggregated locally and no need to send all raw data to the sink to process. The location information can be utilized in routing for network layer. If a network is well-connected, the normal routing protocol can use conjunction with topology control.

### **3.12 Routing Protocol**

Routing protocols are responsible for constructing and monitoring routes between distant nodes. Routing protocols of different characteristics formulate them appropriate for certain distinguished applications. Routing in WSN is very challenging due to constraints which distinguished it from traditional wired network. Routing protocols are classified in many categories:

1. **Data-Centric Protocol:** WSN routing is basically data-centric, not address-centric because there is no need to specify the source of data. Flooding and gossiping [21], SPIN [22], Directed diffusion [23], rumor routing [24], gradient-based routing [25], CADR [26], COUGAR [27], and ACQUIRE [28] are few examples of data-centric routing protocols.
2. **Hierarchical Protocol:** A solitary level system can make the entryway overburden with the expansion in sensor density. Hierarchical protocol efficiency manage the energy consumption by the multi-hop communication within the cluster. LEACH [16], PEGASIS [29], and hierarchical PEGASIS [30], TEEN [31], and APTEEN [32], energy-aware routing for cluster-based sensor network [33], are few examples of hierarchical routing.
3. **Self-Organizing Protocol:** A routing architecture with addressing for each node is proposed and identified by this address to its connected neighboring nodes. The routing architecture is hierarchical. The nodes, themselves, find a route and create a routing table various phases of self-organization.
4. **Location-based Routing Protocol:** Area data are required to ascertain separation between two sensors. There is no addressing scheme like IP addresses, location information can be used to find the route and routing of the data. MECN [34] and SMECN [35], GAF [36], and GEAR [37] are some examples of location-based routing protocol.
5. **Network Flow and QoS-aware Protocol:** In some routing methodologies, courses setup is displayed and tackled as a network flow issue. QoS-aware conventions consider start to finish defer prerequisites while setting up the path in the sensor arrange. Maximum lifetime energy routing, maximum lifetime data gathering, minimum cost forwarding, SAR [17], energy-aware QoS routing protocol [38], and SPEED [39] are key examples of network flow and QoS-aware protocols.

### ***3.13 Topology Control***

In a densely deployed WSN, a single node can establish direct communication with a large number of neighboring nodes of its proximity. Transmission will be increased so that transmission power, which requires lots of energy, increase burden for MAC layer protocol and routing protocol. Instead of highest possible connectivity of the network, a conscious choice can be made to restrict the topology, known as topology control. There are various options for topology control like reduce the set of active nodes/active links and re-arrange the nodes/links in hierarchical network topology.

Common power protocol(COMPOW) and K-NEIGH protocol are key examples of flat networking topology, in which, all nodes are operational and have similar task to perform. LEACH [16] is an example of hierarchical networking by clustering.

### ***3.14 Transport Layer***

In classical layering architecture, it should move free byte streams and middle nodes do not think a lot about the information. Because for intermediate nodes, the packet will be forwarded and processed till network layer only. In a sensor organize, the nodes team up and cooperate with nature, the nodes know the information they convey. For this layer, reliability is a key requirement and coverage of a sensor network is also an important consideration.

There exist various tasks attributed to transport layer, which are reliable data delivery, flow control, congestion control, and network abstraction. Ubiquitous TCP which has designed for traditional transport layer is quite different with respect to WSNs transport layer environment. Some challenges such as multi-hop networking of homogeneous nodes, energy constraints, memory constraints, or computational constraints make it tedious to run heavyweight protocol like TCP.

ReInForM, HHB, and HHBA protocols are some protocols used for providing alternative routes for multi-paths. PSFQ and RMST are block delivery protocol for sink-to-source and source-to-sink, respectively.

### ***3.15 Upper Layers and Advanced Application-Based Support***

The essential functionalities of the layers and protocols which are required to function the WSN properly are described so far. There exist some advance level additional techniques to support for particular task. Advance level data aggregation, distributed signal processing, distributed source coding, and network coding are few in-network processing, adopt by nodes, router, or network to specific task. Some application-specific supports are target detection and tracking, contour and edge detection, and field sampling. Most important concept of additional application support is security.

### 3.16 Security

For many applications of WSN, security is an important requirement. However, security requirements of WSN are same as traditional networks, but the solutions are different. Many existing WSN implementations do not address security requirements. That is a drawback of WSN network protocol. WSN sensor nodes are constrained by memory, bandwidth, and power requirements. Which makes it difficult to deploy complex security algorithms and store security keys of other nodes in any sensor node. So that network designer has to decide to include one or more security goals: Confidentiality, Authentication, Integrity, Accountability, and Availability.

As a threat to WSNs, some of the common attacks are eavesdropping, a passive attack, and insertion, deletion, modification or replaying the packet are active attacks. Distinct types of attacks can be taken place at all the layers of protocol stack.

The countermeasures against these threats are symmetric or asymmetric cryptographic algorithms. Encryption/decryption, Message Authentication Code(MAC), Key Management are some cryptographic algorithm and miscellaneous security protocols are designed on the top of these algorithms.

## 4 Conclusion and Future Work

By this survey, it can be concluded that wireless sensor networks are different from traditional network. This is due to different in the requirements, application-orientation of wireless sensor network, resource constraints, size limitation, and so on. Throughout the survey of various protocol for distinct layers, it can be observed that the prime concern to design a protocol for wireless sensor network is energy efficiency. For future purpose, build such protocol which can emerge more than two specific tasks in one protocol in energy-efficient way. So that the single module of protocols can be used one fit for all fashion with energy-efficient way like traditional protocol generally do.

## References

1. Intel StrongARM SA-1100 Microprocessor Brief Data Sheet. Intel product documentation, August 2000
2. MSP430x1xx Family User's Guide. Texas instruments product documentation, 2004
3. ATmega 128(L) Preliminary Complete. Atmel product documentation, 2004
4. R. F. Monolithics. TR1000 916.50 MHz Hybrid Transceiver, 2000
5. Hill J, Culler D (2002) MICA: a wireless platform for deeply embedded networks. IEEE Micro 22(6):12–24
6. CC2420 2.4 GHz IEEE 802.15.4/Zigbee RF Transceiver. Chipcon Product Data Sheet
7. Wireless Components ASK/FSK 868 MHz Wireless Transceiver TDA 5250 D2 Version 1.6 Infineon Product Data Sheet, July, 2002

8. R. F. Ember—Embedded. Design of an IEEE 802.15.4 Compliant, EmberNet ready and Zigbee ready communication module using the EM2420 RF Transceiver, 2004
9. National Semiconductor LMX 3162—Single Chip Radio Transceiver, 2000
10. <http://wins.rsc.rockwell.com/>
11. Estrin D, Govindan R, Heidemann J, Kumar S (1999) Next century challenges: a scalable coordination in sensor networks. In: Proceedings of the 5th annual ACM/IEEE international conference on mobile computing and networking (MobiCom'99), USC/Information Sciences Institute, pp 263–270, August 1999
12. Romer K, Kasten O, Mattern F (2002) Middleware challenges for wireless sensor networks. ACM Mobile Commun Commun Rev 6(2):59–61
13. Chandra A, Gummalla V, Limb JO (2000) Wireless medium access control protocols. IEEE Commun Surv Tutorials 3(2):2–15. <http://www.comsoc.org/pubs/surveys>
14. Ye W, Heidemann J, Estrin D (2002) An energy-efficient mac protocol for wireless sensor networks. In: Proceedings of INFOCOM 2002, June 2002. IEEE Press, New York
15. Raghavendra CS, Singh S (2000) PAMAS—power aware multi-access protocol with signaling for ad hoc networks. ACM Comput Commun Rev 27:5–26
16. Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Hawaii international conference on system sciences, pp 174–185, Hawaii, HI, January 2000
17. Sohrabi K et al (2000) Protocols for self-organization of a wireless sensor network. IEEE Pers Commun 7(5):1627
18. Rajendran V, Obraczka K, Garcia-Luna-Aceves JJ (2003) Energy-Efficient, collision-free medium access control for wireless sensor networks. In: Proceedings of ACM SenSys 03, Los Angeles, CA, November 2003
19. Greunen JV, Rabey J (2003) Lightweight time synchronization for sensor networks. In: Proceedings of the 2nd ACM international workshop on wireless sensor networks and applications (WSNA), San Diego, CA, September 2003
20. Ganeriwal S, Kumar R, Srivastava MB (2003) Timing-sync protocol for wireless sensor networks. In: Proceedings of the 1st ACM international conference on embedded networked sensor system (SenSys), pp 138–149, Los Angeles, CA, November 2003
21. Hedetniemi S, Liestman A (1988) A survey of gossiping and broadcasting in communication networks. Networks 18(4):319–349
22. Kulik J, Rabiner W, Balakrishnan H (1999) Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings of the 5th annual international conference on mobile computing and networking (MobiCom'99), Seattle Washington USA, Massachusetts Institute of Technology, Cambridge, MA, August 1999
23. Intanagonwiwat C, Govindan R, Estrin D (2000) Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: Proceedings of the 6th annual international conference on mobile computing and networking (MobiCom'00), USC/Information Sciences Institute, pp 56–67, August 2000
24. Braginsky D, Estrin D (2002) Rumor routing algorithm for sensor networks. In: Proceedings of the first workshop on sensor networks and applications (WSNA), Atlanta, GA, October 2002
25. Schurgers C, Srivastava MB (2001) Energy efficient routing in wireless sensor networks. In: The MILCOM proceedings on communications for network-centric operations: creating the information force, McLean, VA
26. Chu M, Haussecker H, Zhao F (2002) Scalable information-driven sensor querying and routing for ad hoc heterogeneous sensor networks. Int J High Perform Comput Appl 16(3):293–313
27. Yao Y, Gehrke J (2002) The COUGAR approach to in-network query processing in sensor networks. In: SIGMOD Record, September 2002
28. Sadagopan N et al (2003) The ACQUIRE mechanism for efficient querying in sensor networks. In: Proceedings of the first international workshop on sensor network protocol and applications, Anchorage, AK, May 2003
29. Lindsey S, Raghavendra CS (2002) PEGASIS: power efficient gathering in sensor information systems. In: Proceedings of the IEEE aerospace conference, Big Sky, Montana, March 2002

30. Lindsey S, Raghavendra CS, Sivalingam K (2001) Data gathering in sensor networks using the energy\*delay metric. In: Proceedings of the IPDPS workshop on issues in wireless networks and mobile computing, San Francisco, CA, April 2001
31. Manjeshwar A, Agrawal DP (2001) TEEN: a protocol for enhanced efficiency in wireless sensor networks. In: Proceedings of the 1st international workshop on parallel and distributed computing issues in wireless networks and mobile computing, San Francisco, CA, April 2001
32. Manjeshwar A, Agrawal DP (2002) APTEEN: a hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks. In: Proceedings of the 2nd international workshop on parallel and distributed computing issues in wireless networks and mobile computing, Ft. Lauderdale, FL, April 2002
33. Shah R, Rabaey J (2002) Energy aware routing for low energy ad hoc sensor networks. In: Proceedings of the IEEE wireless communications and networking conference (WCNC), Orlando, FL, March 2002
34. Rodoplu V, Ming TH (1999) Minimum energy mobile wireless networks. *IEEE J Select Areas Commun* 17(8):13331344
35. Li L, Halpern JY (2001) Minimum energy mobile wireless networks revisited. In: Proceedings of IEEE international conference on communications (ICC01), Helsinki, Finland, June 2001
36. Xu Y, Heidemann J, Estrin D (2001) Geography-informed energy conservation for ad hoc routing. In: Proceedings of the 7th annual ACM/IEEE international conference on mobile computing and networking (MobiCom01), Rome, Italy, July 2001
37. Yu Y, Estrin D, Govindan R (2001) Geographical and energy-aware routing: a recursive data dissemination protocol for wireless sensor networks, UCLA Computer Science Department Technical Report, UCLA-CSD TR-01-0023, May 2001
38. Akkaya K, Younis M (2003) An energy-aware QoS routing protocol for wireless sensor networks. In: Proceedings of the IEEE workshop on mobile and wireless networks (MWN 2003), Providence, RI, May 2003
39. He T et al (2003) SPEED: a stateless protocol for real-time communication in sensor networks. In: Proceedings of international conference on distributed computing systems, Providence, RI, May 2003
40. Wireless Sensor Network. <http://en.wikipedia.org/wiki/WirelessSensorNetworks>
41. MICA2 wireless measurement system. Document Part Number: 6020-0042-04. [www.xbow.com](http://www.xbow.com)
42. Heinzelman W, Kulik J, Balakrishnan H (1999) Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings of the 5th annual ACM/IEEE international conference on mobile computing and networking (MobiCom'99), Seattle, WA, August 1999
43. Younis M, Youssef M, Arisha K (2002) Energy-aware routing in cluster-based sensor networks. In: Proceedings of the 10th IEEE/ACM international symposium on modeling, analysis and simulation of computer and telecommunication systems (MASCOTS2002), Fort Worth, TX, October 2002
44. Chang J-H, Tassiulas L (2000) Maximum lifetime routing in wireless sensor networks. In: Proceedings of the advanced telecommunications and information distribution research program (ATIRP2000), College Park, MD, March 2000
45. Kalpakis K, Dasgupta K, Namjoshi P (2002) Maximum lifetime data gathering and aggregation in wireless sensor networks. In: Proceedings of IEEE international conference on networking (NETWORKS 02), Atlanta, GA, August 2002
46. Ye F et al (2001) A scalable solution to minimum cost forwarding in large scale sensor networks. In: Proceedings of international conference on computer communications and networks (ICCCN), Dallas, TX, October 2001

# A Survey on Routing Protocols for Wireless Sensor Networks



Anita Chandel, Vikram Singh Chouhan, and Sunil Sharma

**Abstract** In recent advances, wireless sensor network energy efficiency is the prime consideration. The routing protocols in WSNs are different from traditional network, application-oriented and depending on network architecture. This paper consists of a survey in the area of routing protocol for WSNs. The routing is categorized into five major sections: data-centric, self-organizing, hierarchical-oriented, location-aware and network flow and QoS-aware. Various protocols fall in these categories described in appropriate section.

**Keywords** Wireless sensor networks · WSN · Nodes · Base station · Routing · Protocol · Resource constraints · Source · Destination · Routing table · Algorithm · Energy-Aware

## 1 Introduction

A wireless sensor network (WSN) is a wireless network, which contains distributed autonomous nodes (sensors). Nodes are capable of capturing information, processing and communicating to each other and routing data back to the base station. Base station issues commands to manage sensors, collects sensors data and processes it. WSN provides a simple and economic approach for the deployment of distributed sensing nodes and controlling them. WSN plays an important role in a wide variety of area ranging from sensing physical and environmental conditions, measuring medical

---

A. Chandel (✉) · V. S. Chouhan

Information Technology Department, Engineering College, Bikaner, Bikaner, India  
e-mail: [anita2506@gmail.com](mailto:anita2506@gmail.com)

V. S. Chouhan

e-mail: [vikksecb@gmail.com](mailto:vikksecb@gmail.com)

S. Sharma

Computer Science Department, Government Polytechnic College, Jhalawar, Jhalawar, India  
e-mail: [sunil13982@gmail.com](mailto:sunil13982@gmail.com)

parameters for health applications, to security-related application as in hostile environment of a battlefield, building security, fire fighting, etc. A WSN consists of many nodes which are either fixed location-based sensors or movable sensors in the monitoring environment.

## 2 Routing Protocols

In a multi-hop network, the role of transitional nodes is to disseminate the packets from source to destination. In similar multi-hop networks, an intermediate node including the source node requires a mechanism to determine which neighboring node would be used to pass the packet to reach at the destination, eventually. The process of passing the packets to next node is known as forwarding.

There exist several different methods to organize the forwarding process. The simplest method, to send a packet to all its neighboring nodes, for the network which contains both the source and the destination in it, is known as flooding. Another method to forward the packet to any arbitrary node, to eventually send the packet to all the neighboring node, is known as gossiping. But both the methods lie on the extreme ends, so there exists an intermediate method called controlled flooding. In this method, the source sends the packet on an arbitrary path or each particular node forwards the packet to a portion of a set of its neighbors.

The performance of the forwarding is measured by the means of number of packets forwarded or delayed. The suitability of a neighboring node completely depends on the cost induced in sending a packet from the source to intended destination using this specific node. The cost can be estimated by number of hops or the energy required, from source to destination. Each node collects the value of the cost for the next hop and arranges these values in a tabular form, called routing table. Various routing algorithms are derived to determine routing tables, and these routing algorithms are the part of routing protocols.

In wired or traditional networks, distance–vector algorithms, like Dijkstra and Bellman–Ford, are used to produce a route. But in multi-hop wireless networks, due to resource constraints, different approaches are prescribed. The properties of a routing protocol for WSNs are low overhead, distributed, self-organizing and handling frequently changed topology.

### 2.1 Network and System Architecture

Wireless sensor network is foremost depending upon the application, hence leads to difference in network requirements, architecture, design goals, resource constraints and computational and communicational usage. It further affects the performance of any routing protocol.

### **2.1.1 Node's Architecture**

In preliminary research in the field of WSN, for all the sensor nodes, an assumption is imposed, to be homogeneous in the terms of communication, computational, node's composition and power. Due to different types of applications and its functionalities, some nodes are quickly drained in similar time duration and lead to the concept of heterogeneous nodes.

### **2.1.2 Node Deployment**

Node deployment is a rudimentary issue in wireless sensor networks. An appropriate deployment technique can decrease the complexity in the terms of routing, data aggregation, computations, communications and power. In large sensor networks, the sensors are deployed by aerial means and creating a random infrastructure. The nodes have to be self-organized and able to locate the neighboring nodes.

### **2.1.3 Energy Considerations**

Energy consideration is the prime factor during the formation of an infrastructure. The energy consumption of the wireless radio is proportionate to the square of distance or presence of obstacles of higher order. Direct communication is performed well for the nodes nearer to sink, and multi-hop consumes less energy for larger distance. However, multi-hop introduces the notable expenses for topology management and MAC.

### **2.1.4 Data Delivery**

The data conveyance to the sink in WSNs can be categorized into persistent, event-based, query-based and hybrid based on the requisitions of the networks. The data delivery scheme is highly influenced by the minimization of energy dissipation and route sustainability.

### **2.1.5 Data Aggregation**

Since each node senses and forwards the sensed readings to the sink, it generates a large number of redundant packets. Data aggregation combines the information from different resources by applying different functions. This technique is utilized to reduce battery consumption and congestion optimization in various routing protocols.

### 2.1.6 Network Dynamics

Most of the network setup is assumed to be stationary, and the topology is fixed. Routing in static network is less challenging. On the other hand, some network is completely mobile and topology is changed according to the time. Tracking the required node is also added with the basic functionality of routing.

## 2.2 *Characteristics*

There are several characteristics that make routing in WSN very challenging and distinguish from traditional networks.

1. **No Global Addressing:** Due to preparation within the unattended and remote field, it is impracticable to create global addressing strategy. Therefore, classical IP-based protocols cannot be applied to WSNs.
2. **Multiple Source–Single Destination:** Unlike, one to at least one traditional networks, the majority applications of WSNs need to produce the perceived information from multiple sources to single destination.
3. **Redundancy:** Multiple sources may produce the same data within the surroundings of a WSN. It will be further responsible to produce redundant data. This redundancy is taken care by the routing protocols to conserve the energy and bandwidth employment.
4. **Resource Management:** Motes are highly resource restrained in the terms of communicational power, battery, processing and storage capacity, so the available resources should be managed appropriately.

## 2.3 *Classification Based on Number of Recipients*

1. **Unicast:** In the unicast, number of sender and receiver is strictly one each. Unicast is preferably simple: consider a network graph, assign the load to each edge which is the energy consumption for that link and apply any shortest path algorithm to compute the minimum cost of the network. Unicast is sub-classified into single path unicast and multi-path unicast.
2. **Multicast and Broadcast:** In the broadcast operation, the information is collected and forwarded to all the nodes in network. This operation is so frequent in WSNs that the basic nature of WSN is assumed to be broadcast and unicast is application-oriented. When some data are distributed to a given subgroup of the nodes of the particular network, it is known as multicast.

### 3 Data-Centric Protocol

In a giant sensing element network, it is not possible to assign global identifiers to every node and in conjunction with random preparation makes it onerous to pick a selected set of sensing element nodes to be examined. This initiates to adopt different approaches from traditional routing, in which the route is managed in addressable nodes.

In data-centric routing, the sink sends the query to the network and waits for the perceived information from the nodes from its proximity. Attribution-based naming is needed to particularize the characteristics of data.

#### 3.1 *Flooding*

Both play an important role in the design of communication protocols in various kinds of networks, to transmit the data without routing algorithm and topological control. The gossiping problem can be defined as: in a set of numerous individuals, all have an item of information to gossip that they want to forward to everyone else. Communication is generally done in a set of rounds, and in each round, an individual may communicate to at most one other node. The communication can be represented as a graph, where each edge shows the link between pairs of nodes one at each round.

#### 3.2 *Gossiping*

The broadcast or flooding problem can be defined as: one individual node has to convey an item of information to every other individual node in the network. Total number of communication rounds and links between the sensors are two basic parameters, which are used to evaluate the algorithms for broadcasting or flooding problem. However, it is very easy to implement, and it has some serious drawbacks: implosion, overlap and resource blindness [1].

#### 3.3 *SPIN*

Sensor Protocols for Information via Negotiation (SPIN) [2] is a family of accommodative protocols that with efficiency distributes data among the nodes in energy-efficient approach in wireless detector network. The efficient propagation of the information through individual node to neighboring sensors of the network is observed, and each and every sensor is treated as potential sink nodes. This approach has two

benefits incomplete. First, it gives numerous replicating views of the network to improve the fault tolerance and, second, spread critical information to all the nodes.

A conventional protocol, classical flooding is compared and analyzed as the base of this approach. Five different protocols are evaluated by simulation for efficient information dissemination, but two among them are experimental protocols: SPIN-1 and SPIN-2. SPIN-1 employs arrangements to resolve the problem of inadequacy and overlap. SPIN-2 utilizes a threshold-based resource-aware mechanism with negotiation to solve the problem of resource blindness.

The basic plan is to call the information by victimization of high-level descriptor or metadata. However, SPIN does not specify a format for metadata. So as to conserve the energy, detector applications communicate with one another regarding the information that is sent or about to send. Sending the data is expensive but not the data about the data. Second, node must monitor the change to their energy resource. Metadata eliminates the possibility of redundant data by allowing the node to place interest of particular portion of data. The data will be sent only when it is required and reduces waste of energy in useless transmissions.

There are three messages defined in SPIN: ADV: to advertise the particular metadata, REQ: to request the specific data and DATA: to carry the actual data. SPIN is an application-based approach.

### 3.3.1 SPIN-1

It is a simple handshaking protocol for a lossless network. The protocol starts when new data is obtained by any node. The node sends ADV message to its neighboring nodes. After receiving the ADV, neighboring node checks that it is requested or already received. If not received, then it will send a REQ message back to the sender and get DATA in response. The main drawback of this protocol is that it is designed for lossless networks only.

### 3.3.2 SPIN-2

The SPIN-2 protocol is similar to SPIN-1 with addition to a low-energy threshold. Whenever the energy is leading to the minimum energy bound, it remodels to reduce its contribution to the network activities.

## 3.4 *Directed Diffusion*

Directed diffusion [3] is an application-aware data-centric approach in which the communication is performed on the basis of named data. By choosing a decent ways and caching and process knowledge inside the network, it ends up in delivering the goods energy conservation. Attribute–value pair is used to name the data. A list of

attribute–values pairs describe a task, and these task descriptors are named in this approach. A sensing task is broadcasted all over the network as an interest by a node called a sink. The sink used to broadcast the interest message to the network, periodically. The interest is periodically refreshed by the sink. An interest table is organized and maintained by every node, which has each distinct entry.

When an interest is received by a node, then it checks either the interest is already present or not. If not present, then an entry is created for the interest and a single gradient toward the particular neighboring node is also added. When the gradient is timed out, the interest entry would be removed. After receiving a fresh interest, the node may broadcast the same interest to other neighboring nodes also. A gradient specifies a data rate and a direction of the send event. The interest flew in such a manner to facilitate the data toward the sink node. The data packet is unicasted to the relevant node or neighboring node only. A matched cache entry is dropped eventually.

Directed diffusion has numerous characteristics like no router required, no global identifiers required and reactive routing.

### ***3.5 Gradient-Based Routing***

The Gradient-Based Routing (GBR) is a variation of the directed diffusion (DD) convention. It is a question-based steering convention in which information is assembled and sent to the base station by sensors or a gathering of sensors, in a response of a query, generated and broadcasted by a sink node. The query is circulated in the form of interest message, which contains description of the task, from a sink node.

There are two approaches which are proposed to, first, aggregate packet stream in a vigorous way, to minimize the energy dissipation by a number and, second, to get better resource utilization in order to achieve longer lifetime for sensor network. Like SPIN, when new node is entered in the sensor network, it announces the type of interest by broadcasting, and a gradient will be generated at each sensor. The gradient indicates the integrity of the nodes which are few hops away and used in packet transmission. The height of node is calculated by minimum number to the interest. The difference between the height of the node and the height of neighboring node is known as gradient. The packet will be sent on the link which has the largest gradient. Each node processed their data before forwarding to the next hop. The concept of Data Combining Entity (DCE) is introduced for the node which have multiple streams, flowing through them. It compacts the data into one unit for transmission.

### ***3.6 Energy-Aware Routing***

The key concern of the protocol, energy-aware routing [4] is, always considering most minimal vitality path may not be ideal for the system lifetime and long haul network

viewpoint. For optimizing such measures, sub-optimal paths may be established for substantial gain. By analyzing few existing energy optimal protocol lead to discover the optimal path, burn a lot of energy along with these paths and leave the path with huge disparity in energy level and eventually few nodes die; and disconnect the network. The protocol did not search for the single optimal favorable path. Rather a set of sub-optimal paths is discovered, and one path is chosen in probabilistic way.

The primary idea of the protocol is to improve the lifespan of the network that may lead to use sub-optimal paths sometimes and ensures that it does not reduce the performance of the network. None of the paths used all the time in order to maximize sustainability of the network and instead of that use different paths continuously. The protocol is a reactive type of routing protocol, hence destination initiated. It is similar to directed diffusion and compares the performance with the protocol.

The protocol works in three phases:

1. **Setup Phase or Interest Propagation:** Destination initiates the connection by flooding the request, and cost is initialized to zero. Every intermediate node forwards the request, and cost would be added to the previous cost. At the destination, path with high cost is discarded and low-cost paths are added to forwarding table. An average cost is calculated for each node.
2. **Data Communication Phase:** The source and intermediate routers forward the information to the neighbors in forwarding table with the neighbor's likelihood equivalent to the normal likelihood of the node's forwarding table. The information sent till the destination.
3. **Route Maintenance Phase:** The restricted interest flooding performed sometimes, and subsequently, course support is low.

The simulated results show that energy-aware routing reduces the energy consumption at each node and has an improvement of 21.5% in performance.

### **3.7 Rumor Routing**

Rumor routing [5] is a method to describe and analyze the routing queries of the sensor nodes that are continuously observing a particular event and retrieve the data. In more appropriate description, rumor routing is a legitimate trade-off between flooding inquiries and flooding event notifications.

The idea is to establish a path toward the event, whereas the event flooded by broadcasting. At whatever point a query is produced, it can continue to an arbitrary path till the event way is found, rather than communicating the question all through the system. As the query finds the event path, the event straightforwardly courses to the event. In the event that the path is not found, then the query either re-submitted or flooded it. In various tested conditions, an extremely high delivery rate is achieved. The rumor routing is an algorithm to occupy the space between algorithms called query flooding and event flooding. When no prior information about the vicinity is

available, then the query is flooded to the entire network, which is known as query flooding. When a sensor encounters an event, it floods it to the network.

The rumor routing can be described as: (1) Each node keeps a list of neighboring nodes and event table. (2) When a node encounters an event, it will add it into event table with zero distance. (3) An agent is a packet with extended lifetime, transmitting the information about the local event to the distant nodes. (4) All the nodes are capable to generate the query and routed to the particular event. (5) If the query is somehow not able to reach the destination, then the query is either re-submitted or flooded to all the network.

The rumor routing protocol provides an efficient and powerful algorithm to deliver the queries to the particular events in numerous conditions in a large network.

### 3.8 CADR

The idea of constrained anisotropic diffusion routing (CADR) [6] is a part of two techniques that work together, named information-driven sensor querying (IDSQ), for making data querying and routing more energy efficient. The principle is to introduce a mechanism to select a sensor to query and process the data routing, augment data gain, limit discovery idleness and limit the energy utilization by tasks as localization and tracking.

It utilizes the general type of data utility that determines the data content and spatial configuration of the network to formulate how each node estimates the cost details, comes to any decision, updates belief status and routes the data based on various decisions. CADR is a generalized form of directed diffusion routing. The sensors are activated once the particular interest event is announced, and only for the part of the network, that seeks the information.

To formulate the problem, sensing observation model is designed for nonlinear relations between the sensor nodes, types, position, noise and other parameters and measures the uncertainty. Then, consider the sensor node selection by updating the belief by incorporating based on previously not considered nodes. After that characterize the measure for data utility, for example, co-difference based, Fischer data grid, entropy of estimation vulnerability, volume of high likelihood area and sensor geometry-based measures. Then, a composite objective function is derived to calculate transmission cost of the network and current belief. All the routing decisions are based on local computation on each node to achieve the estimated belief status.

### 3.9 COUGAR

COUGAR [7] uses the idea of declarative query, especially designed for sensor networks, to abstract the query processing from other network layer functionalities, by additional query layer. So as to diminish asset energy and, subsequently, increment

the lifetime of the sensor, a query optimizer is utilized to structure a proficient inquiry plan. General existing networks collect and transfer the data when they are online and aggregate and store for offline querying and analysis. It also proposed a loosely coupled distributed architecture to support data aggregate and complex computations.

A query optimizer is also situated on the gateway node to design a plan for the processing of distributed query whenever it receives the query. The query plan architects the flow of data between the nodes and specifies the computation at each node. The plan is broadcasted to all the sensor nodes in the network. At query execution time, data record is sent back to the gateway node.

The protocol works in various phases: (a) data aggregation, (b) selection of query language, (c) optimizing the query, (d) catalog management and (e) optimizing the multi-query.

### **3.10 ACQUIRE**

ACtive QUery forwarding In sensoR nEtworK (ACQUIRE) [8] considers the system as circulated database and works with complex inquiries which are additionally partitioned into sub-queries. The standard behind ACQUIRE is to infuse a functioning query bundle that finishes a path of the system. At each progression, the node, which gets the dynamic query plays out an activated, on-request, redesign getting data from every one of the neighbors within the scope of countable hops.

As this dynamic query advances through the system, it gets progressively separated into littler segments till it is totally comprehended and returned back to the underlying query creating node with a total outcome. The basic concept is based on test and validation of data query, but, on the other hand, a mathematical approach is also proposed to derive analytic expressions for the energy cost of complete ACQUIRE protocol. The efficiency of ACQUIRE can be enhanced if the surrounding nodes of the active nodes in the query path have the minimum intersection and additional topological or geographical information which helps to perform the task.

In this mechanism, the query is forwarded to the nodes in the network by the sink and each node, which gets the query, attempts to process the query completely or incompletely by utilizing pre-reserved data and forwards to the following sensor node.

## **4 Self-organizing**

In WSNs, various simple components demonstrate the behavior of the whole network and represent more arranged than the behavior of individual component. Subramanian and Katz [9] proposed a collective architecture for specific sub-class of sensor applications, named self-configurable system, for large number of sensor to co-ordinate themselves to complete a sensing task. The main goal of the algorithm is to

minimize energy consumption, localizing operations, fault tolerance for node and link failure. The network applications can be classified based on size of the system, numbers of sensors used, distance between the nodes from wired component and allocation of sensor nodes.

There are few assumptions built for generic architecture of self-configurable system, such as nodes should be heterogeneous, data delivery and data dissemination are two different and orthogonal components, resource constraints and infrastructure requirements are application-specific, and data discovery nodes are mobile, whereas data dissemination nodes are stationary. However, in the presence of heterogeneous node, all nodes should be treated the same and have similar functionalities.

The architecture supports both data-centric and non-data-centric approaches. Reduction of state and localized operations, energy efficiency, reliable paths, hierarchical routing paths, fault tolerance using broadcasting tree, and reduction of the number of dynamic cost updates are few contributions of the algorithm.

## **4.1 Architecture**

In the architecture, the nodes are assumed to be stationary. Various architectural components are classified such as specialized nodes to monitor and track; router nodes for information dissemination and strictly stationary; aggregator nodes to collect and aggregate sensor readings and sink node for high capacity of information storage. Numerous infrastructures are also described like addressing infrastructure; however, individual node does not need any addressing technique but unique MAC address can be used if necessary, routing infrastructure to establish the links between adjacent router nodes, and broadcasting and multicasting infrastructure.

## **4.2 Algorithm**

The self-organizing algorithm can be explained as:

1. **Discovery Phase:** Discovers set of neighboring nodes and fixing the maximum radius of data transmission.
2. **Organizational Phase:** Numerous operations are performed like nodes form the groups and balance the hierarchy, and node's position in hierarchy, routing table is computed by  $O(\log n)$ , and broadcasting trees are constructed.
3. **Maintenance Phase:** Monitoring, constant updation of routing table, details of neighboring nodes, and fault tolerance are performed in maintenance phase.
4. **Self-Reorganization Phase:** As changes occur in topology, node failure, group partitioning or routing table updation, self-reorganization is performed for remedial task and maintains the hierarchy.

### 4.3 Analysis

There is a list of advantages of self-organization like hierarchy is balanced, routing state is maintained by just  $O(\log n)$ , node failure and link failure can be tolerated by Local Markov Loops, broadcasting graphs using directed acyclic graph in a unique manner and attachment to specialized node makes the node mobile. However, there are few disadvantages also such as no on-demand organization, hierarchical operation increases the probability of reorganization and the last but the most important, no address is attached for communication from one node to another.

## 5 Hierarchical Protocols

In a single-tier network, the overhead increases with the sensor density. Hierarchical routing is proposed to productively keep up the energy utilization by including the idea of multi-hop correspondence in the scope of specific group.

### 5.1 LEACH

Low Energy Adaptive Clustering Hierarchy (LEACH) [10] is a self-organized, adaptive clustering protocol in which the distribution of energy load is even among all the sensor nodes. In LEACH organization, the nodes arranged themselves in a local cluster and one node is designated as local base station or cluster head. Cluster head is chosen in rotational basis, because fix cluster head like conventional clustering algorithm leads to drain one node quickly; hence, energy drainage is monochromatic all over the network.

Local data fusion is applied to the data, sent from cluster to base station, in order to minimize the energy consumption. Each node chooses the cluster head which is close in vicinity; hence, minimum communication energy will be required. Once all the nodes are arranged in cluster-based organization, a schedule is prepared for the node in a particular cluster, to turn off all the non-cluster head node at the time of transmission. Collected data are aggregated at cluster head and transmitted to the base station. The node with maximum energy remaining is chosen to be cluster head, so that distant base station transmissions can be easily performed.

The operation of LEACH is first divided into rounds, and these rounds are further divided into various phases.

1. Advertisement Phase: The node that elected itself as a cluster head advertises itself by broadcasting a message to rest of the nodes.
2. Cluster Setup Phase: Once cluster head is selected, the nodes have to choose their belonging to the cluster and inform the cluster head about their membership.

3. Schedule Creation: The cluster head creates a TDMA schedule for the node to transmit.
4. Data Transmission.

The main advantage of LEACH is completely distributed and requires no control information from the base station, and the node does not need any prior information of the global network.

## 5.2 PEGASIS

Power-Efficient GAthering in Sensor Information System [11] (PEGASIS) is a near-optimal chain-based protocol and built on the top of LEACH. The fundamental thought for PEGASIS is to shape a chain of sensors so that every node can transmit and get from the nearby neighbors as it were. Gathered data forward starting with one sensor then ont the next and get intertwined, and at last, a chosen node transmits it to the base station.

Nodes take turn in rotation basis, to transmit the data to base station, in order to reduce the average energy consumption at each node per round. Intractable problem like traveling salesman problem has resemblance to build a chain to minimize the total length, hence adopts greedy approach. In the case of no node in transmission range or no node left out, multi-hop organization can be used. In any case, the node died at random place, a simple control token passing approach is actuated to start the data transmission from the end of the chain. Local gathering and small amount of data received by the leader are two prime improvements of PEGASIS over LEACH to save energy for certain levels.

## 5.3 Hierarchical PEGASIS

Hierarchical PEGASIS [12] is an extension of PEGASIS. It evaluated the data gathering problem, in the case, where data is gathered in each round and combined with other nodes' data, encapsulated in one unit and forwarded to the farther base station. On the off chance that information is sent over unit delay and transmitted straightforwardly to the base station, at that point both energy utilization and postponement would be high. The thought is to build up an information-gathering component that equalizes the vitality and defer cost, as measured by energy \* delay. A chained-based binary scheme is best suited with CDMA capable nodes in terms of energy \* delay. On account of without CDMA fit nodes, parallel interchanges are conceivable just among spatially isolated nodes and a chain-based 3 level progressive system plot fits best for the plan.

## 5.4 TEEN

Threshold-sensitive Energy Efficient sensor Network (TEEN) [13] is an energy-efficient protocol for active network and evaluated on simple temperature sensing application. In this approach, the network is classified on the basis of their mode of functioning and the type of target application, as proactive network and reactive network. In proactive, the network is switched on the sensor and transmitter, periodically, senses the environment and transmits the sense information. Whereas, in reactive, the node reacts to the changes based on sensed attributes. The model uses a hierarchical clustering approach.

All the nodes transmit the data to their immediate cluster head only, to conserve the energy, and the cluster head is intended to perform additional computations. Cluster head would be changed periodically, based on the battery left. The cluster head broadcasts the parameters, which are report time and attributes of network. Report time is the time to sense the information and transmit the data. Time-critical data reaches the node without any time lost. Soft threshold may vary, but smaller soft threshold value shows a broader image of the whole network. Whenever cluster changes time, the attributes broadcast again. The main drawback of this scheme is if the thresholds are not reached, the nodes will never communicate, and the whole network dies without any transmission. Hence, it is not suitable for the applications where the data is required at regular basis.

## 5.5 APTEEN

Adaptive Periodic Threshold-sensitive Energy Efficient sensor Network (APTEEN) [14] is a hybrid routing protocol which is designed for extensive information retrieval. Like LEACH and TEEN, APTEEN also chooses a cluster head (CH), and once the CHs are decided, in each cluster report time, the cluster head first broadcasts the parameters, such as attributes, threshold (soft and hard threshold), schedule and count time ( $T_c$ ). TDMA is introduced to avoid the collisions between the data packets. Time-critical situations are also arising in the network due to various changes. The protocol is able to handle three types of queries: historical query, one-time query and persistent query. These queries collect the data on periodic base with hard threshold. The main drawback of this scheme is additional complexity of thresholds and time counts. APTEEN is an extension for WSN with uneven node distribution.

## 6 Location-Based Protocols

Location is the prime concern in WSN when the distance between two nodes is to be calculated so that the energy dissipation can be calculated. On account of no tending to conspire, area data can be utilized in steering in energy proficient way.

### 6.1 SPEED

SPEED [15] is a real-time, stateless, localized algorithm with minimal control overhead communication protocol. Location information with data makes location-based routing more meaningful for decision making. Unlike other location-aware protocols, SPEED is able to handle congestion control with soft real-time communication service. MAC layer and network layer adaption are used to deal with such issues. SPEED works to achieve a soft real-time communication service on a certain delivery rate and the delay between the end nodes is proportional to the distance between them. Few design objectives are mentioned to be satisfied by SPEED are architecture should be stateless, soft real-time, minimal support by MAC layer, congestion management, traffic and load balancing, localized behavior, ineffective path avoidance and no routing table only location information.

SPEED achieves a desired delivery speed throughout the network by diverting the traffic and regulating the packets locally. SPEED contains various components like application API, packet format, neighboring beacon exchange mechanism, estimated the delay, stateless non-deterministic geographic forwarding, neighborhood feedback loop, back-pressure re-routing and the last mile process. SPEED utilizes non-deterministic sending to accomplish balance in energy utilization. It actualized on Berkeley motes platform whose code size is 6036 bytes.

### 6.2 MECN

Minimum Energy Communication Network (MECN) [16] is a convention for disseminated position-based system to improve the base energy utilization in versatile remote system for distributed correspondences. To accomplish such correspondence, the system must be firmly associated. The most widely recognized channel model utilized for RF frameworks and have different parts like way misfortune, enormous scale varieties and little scale varieties. Maximum ratio combining is the method used to join these streams ideally. In the path-loss model, the statures of antenna of mobile ought to be comparative.

In order to establish a strongly connected network, each node keeps communication links to every other node. By using a relay node, minimize the total energy consumption between sender and receiver nodes. The key plan to the protocol has

to consider every one of the nodes of the system to locate the worldwide energy proficient way to the ace site by every sensor. By local search, it considered only few potential links in the relay region. The protocol works in two phases: (1) Search for Enclosure: To discover a enclosure graph, neighboring node and enclosures are searched by every node and no loops are considered in graph. (2) Cost Distribution: It finds the optimal links with minimum power dissipation in the enclosure graph. The relay region registered for a solitary node by accepting the two-ray proliferation model for area correspondence. This is applicable for only mobile ad hoc network, because the topology is discovered by local search performed on neighborhood of each node.

### 6.3 SMECN

Based on the minimum energy property, here present the characteristics to establish a protocol named Small Minimum Energy Communication Network (SMECN) [17] on the top of MECN protocol. A sub-network is constructed by SMECN and proved that the broadcast would be reached to all the nodes in a circular region, if happened at given power setting, is smaller than MECN. The key idea to minimize the energy consumption is to construct a subgraph  $G'$  of original network graph  $G$  such that

- $G'$  consists of all the nodes of  $G$  with fewer (necessary) edges.
- If two nodes are connected in  $G$ , then it must be connected in  $G'$ .
- Each node in  $G'$  can transmit to all neighboring node by consuming less energy than  $G$ .

The idea is to propose an algorithm to construct a communication network which contains  $G$  min rather construct  $G_{\min}$  itself. To find the routes for message sending, it adopts AODV approach.

### 6.4 GAF

Geographic Adaptive Fidelity [18] protocol is introduced to minimize energy consumption in wireless ad hoc network by discovering route equivalent sensor node and turned off unwanted node in order to maintain routing fidelity. For this purpose, the sink and source node are remained constant and intermediate nodes are examined and used for balancing the energy consumption. Adaptive fidelity is a technique to increase the lifetime of self-configured system by eliminating the redundancy.

An algorithm introduced for GAF is as follows:

- Determining Node Equivalence: The concept of virtual grid is introduced so that each GAF node can associate with it by using its location information and all the node from the same square of the grid are equivalent.

- GAF State Transitions: Nodes are designed to be in one of the three states: sleeping, discovery and active. The node is initially in discover state, till the timer fired, and then node broadcasts the message and enters in active state for a time stamp. A node can change the state to sleeping when equivalent node will handle routing.
- Tuning GAF: The value of various parameters including estimated node active time (ENAT), discovery message interval ( $T_d$ ), active time duration ( $T_a$ ), node ranking and sleeping duration ( $T_s$ ) is application-specific in GAF.
- Load Balancing Energy Usage: GAF introduces a load balancing strategy to reduce energy uses so that all the nodes remain on and work together for longer lifetime.
- Adapt to Higher Mobility: GAF tries to maintain a constant level of nodes to participate in data routing, but high mobility increases the packet drop rate. Hence, GAF classified the simulation in GAF-basic and GAF-mobility adaption.
- GAF can run with any ad hoc routing protocols and decision of turning on and off for the nodes is completely independent.

The overhead generated by GAF discovery message is small and not affects the energy consumption too much. GAF may affect the data delivery by losing the packet when sleep mode is on and route is changed.

## 6.5 GPSR

Greedy Perimeter Stateless Routing (GPSR) [19] is a protocol introduced for wireless datagram networks, which forward the data by using position of the router and destination, and a greedy strategy is to adapt for immediate nodes in the network topology. In the case of frequent topology change, GPSR uses local topology information to establish accurate new route quickly.

The prime factors to build a routing algorithm are rate of change in topology and number of routers used. Hierarchy and caching are two promising approaches for scaling ad hoc routing. The algorithm is divided into two approaches for packet forwarding: greedy forwarding, in which a greedy choice is obtained in choosing the next hop with local optima, perimeter forwarding, used where greedy forwarding cannot be applied.

A simple beaconing algorithm yields the position of the neighboring node position of all the existing nodes of the network by broadcasting MAC address, identifiers and position in the form of a beacon by every node. When the node is failed to receive any beacon from neighbor node, node assumes that the node is out of range or die and deletes it from routing table. In perimeter forwarding, the next edge to be traverse is the one in counterclockwise to the edge. In the shortage of no crossing heuristic, the idea of planer graph is introduced. If a neighbor is closer to the destination, the packet is forwarded to next hop, else packet is marked as perimeter mode and use a simple planer graph traversal.

For protocol implementation, GPSR has to make more robust on a mobile IEEE 802.11 network, by adequate choice of implementation like reinforcement in the case of MAC layer failure, interface for query traversal, efficient use of network interface and planarized the graph. GPSR is proficient to utilize geographic parameter to achieve small per node routing state, less message complexity and robust data delivery in dense wireless network.

## 6.6 *GEAR*

Geographic and Energy Aware Routing (GEAR) [20] evaluates an energy-efficient routing protocol to broadcast a query to the particular region of the network without flooding. Neighboring node is chosen in an energy-efficient way to route the packet in target region and uses restricted flooding inside destination region to reduce duplicate forwarding. GEAR is simulated for both uniform and non-uniform traffic and compared to the performance of GPSR.

The packet forwarding process comprises two phases to forward packet to all the nodes in target region. First, forwarding toward the target region. If the node is closer, neighbor to the destination sends to next hop, else sends to next hop with minimized cost. Second, diffuse the packet within the target region by recursive geographic forwarding. For that, the targeted region should be specified, node's remaining energy must be known and link should be bi-directional. The energy-aware neighbor is computed for closer and farther away nodes. In recursive geographic forwarding, the recursive splitting and forwarding are followed till the pre-conditions are not fulfilled. The energy-aware metric for cost estimation is used to balance energy consumption and achieve energy efficiency for the network. As the simulated result, it shows that GEAR delivers 70–80% more data packets than GPSR.

## 7 Network Flow and QoS-Aware Protocols

This is not somehow the routing protocol by intended nature, but when the route is modeled in such a way to solve network flow problem. QoS protocols look after the end-to-end delay while setting up the routing paths. There are variety of protocols in this section: Maximum lifetime energy protocol [21], maximum lifetime data gathering [22], minimum cost forwarding [23], sequential assignment routing [24] and energy-aware QoS routing protocols [25] are some examples.

## 7.1 Maximum Lifetime Energy Routing

Maximum Lifetime Energy Routing [21] is basically a data-centric routing, but it provides a certain level of QoS, hence considered in this category. The basic idea is that the system lifetime can be significantly extended by using a link metric that utilizes the information about the residual energy of the sensor nodes and the energy expenditure in the transmission of a unit of information over the wireless link. The approach is further divided into two modules or sub-problems, called the maximum lifetime problem and the maximum residual energy path routing algorithm. Maximum lifetime problem can be described as sensor and can only generate sense packets in detection time only and forward it to one of the gateway nodes. The objective is to amplify the time until the first failure of the packet conveyance because of battery waste. The idea behind maximum residual energy path routing is to route packets through paths with maximum energy remaining so that total energy consumption in all the paths will be balanced. Some of the link metrics and methods of shortest path calculation are compared in simulation part.

## 7.2 Energy and QoS-Aware Routing

An energy-aware QoS routing protocol is proposed for sensor systems for the transmission of imaging information, so as to boost the effectiveness of the sensor and strong methodology for data gathering. The convention acquires a least-cost, delay-compelled way for continuous applications like picture, as far as cost of connection, transmission energy, error rate, battery left and different parameters. Throughput for non-ongoing information can be augmented by administration rate modifications.

The greater part of the QoS put together routing algorithm are based with respect to the idea of portability of the node, yet energy efficient, with QoS parameters are not considered by any means, subsequently, challenges presented by imaging sensors and constant applications are not assessed. The possibility of energy-aware QoS routing is to find an ideal way to the door with less energy and error rate, by taking start to finish postpone necessities in representing continuous information. QoS routing problem is similar to path constrained path optimization (PCPO) problem, which is an NP-complete problem.

A ratio is used to derive the bandwidth for real-time and non-real-time traffic for outgoing link, and this ratio is set by the gateway. There are two different mechanisms, for setting the service rate on each node, named single-r and multi-r. In single-r mechanism, one r value is assigned for each node all over the network. But in multi-r mechanism, the sensor nodes have to calculate their own r value based on the information broadcast by the gateway. The multi-r value provides better end-to-end delay, increases throughput, extends queuing delay and increases average delay per non-real packets with a slight increase in energy consumption.

### 7.3 Maximum Lifetime Data Gathering and Aggregation

The protocol is interested to find an efficient way to collect the data from all the sensor nodes and transmit to base station, in order to maximize the system lifetime. The protocol mainly focuses on the performance objective of interest to maximize the lifetime of system instead of trying to minimize the energy consumption. The maximum lifetime data gathering problem is sub-divided into two baselines: sensors are capable to perform in-network aggregation of data; and aggregated data are not being aggregated by other nodes.

The data gathering problem considers two sub-classes as data gathering with aggregation and without aggregation. Each sensor generates one data packet of size  $n$  bits in a particular time stamp, named round, and transmits it to the base station. All the information from all the sensors in one round, gathered and transmitted to the base station. A data gathering schedule  $S$  is specified to achieve maximum lifetime  $T$  for the system.

Data gathering can be achieved by aggregation and without aggregation. In data gathering with aggregation, the data is aggregated by in-network to a data fusion packet to reduce the number and size of the communication and make it energy efficient. The problem in with aggregation is known as Maximum Lifetime Data Gathering (MLDA) and solved by finding a near-optimal admissible flow network and constructing a schedule for that. For without gathering, the problem is known as Maximum Lifetime Data Routing (MLDR) problem. A near-optimal solution can be obtained by MLDR problem, solves a linear relaxation for defined integer problem and computes a solution for the same. In the case of data gathering without aggregation, schedule generated for MLDR is able to achieve better lifetime than existing protocol.

## 8 Conclusion

In this paper, we have outlined recent research in the field of routing in wireless sensor network, which falls into five categories mentioned above. The prime concern of these protocols is energy efficiency and making sure of participating of each node in the process of data gathering and transmission. In future, the traditional routing method variants can be used in energy-efficient way for wireless sensor network, with less communicational and computational overhead.

## References

1. Hedetniemi S, Liestman A (1988) A survey of gossiping and broadcasting in communication networks. Networks 18(4):319–349

2. Kulik J, Rabiner W, Balakrishnan H (1999) Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings of the fifth international conference on mobile computing and networking (MobiCom'99), Seattle Washington USA, Massachusetts Institute of Technology, Cambridge, MA, August 1999
3. Intanagonwiwat C, Govindan R, Estrin D (2000) Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: Proceedings of the 6th annual international conference on mobile computing and networking (MobiCom'00), USC/Information Sciences Institute, pp 56–67, August 2000
4. Shah R, Rabaey J (2002) Energy aware routing for low energy ad hoc sensor networks. In: Proceedings of the IEEE Wireless communications and networking conference (WCNC), Orlando, FL, March 2002
5. Braginsky D, Estrin D (2002) Rumor routing algorithm for sensor networks. In: Proceedings of the first workshop on sensor networks and applications (WSNA), Atlanta, GA, October 2002
6. Chu M, Haussecker H, Zhao F (2002) Scalable information-driven sensor querying and routing for ad hoc heterogeneous sensor networks. *Int J High Perform Comput Appl* 16(3):293–313
7. Yao Y, Gehrke J (2002) The COUGAR approach to in-network query processing in sensor networks. In: SIGMOD record, September 2002
8. Sadagopan N et al (2003) The ACQUIRE mechanism for efficient querying in sensor networks. In: Proceedings of the first international workshop on sensor network protocol and applications, Anchorage, AK, May 2003
9. Subramanian L, Katz RH (2000) An architecture for building self configurable systems. In: Proceedings of IEEE/ACM workshop on Mobile ad hoc networking and computing, Boston, MA, August 2000
10. Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the Hawaii international conference on system sciences, 4–7 January 2000, Maui, Hawaii. IEEE
11. Lindsey S, Raghavendra CS (2002) PEGASIS: power efficient gathering in sensor information systems. In: Proceedings of the IEEE Aerospace Conference, Big Sky, Montana, March 2002
12. Lindsey S, Raghavendra CS, Sivalingam K (2001) Data gathering in sensor networks using the energy\*delay metric. In: Proceedings of the IPDPS workshop on issues in wireless networks and mobile computing, San Francisco, CA, April 2001
13. Manjeshwar A, Agrawal DP (2001) TEEN: a protocol for enhanced efficiency in wireless sensor networks. In: Proceedings of the 1st International workshop on Parallel and Distributed Computing Issues in Wireless networks and mobile computing, San Francisco, CA, April 2001
14. Manjeshwar A, Agrawal DP (2001) APTEEN: a hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks. In: Proceedings of the 2nd international workshop on Parallel and distributed computing Issues in wireless networks and mobile computing, Ft. Lauderdale, FL, April 2002
15. He T, Stankovic JA, Lu C, Abdelzaher T (2014) SPEED: a stateless protocol for real-time communication in sensor networks. *IEEE Trans Wireless Commun* 12(9)
16. Rodoplu V, Ming TH (1999) Minimum energy mobile wireless networks. *IEEE J Select Areas Commun* 17(8):1333–1344
17. Li L, Halpern JY (2001) Minimum energy mobile wireless networks revisited. In: Proceedings of IEEE international conference on communications (ICC01), Helsinki, Finland, June 2001
18. Xu Y, Heidemann J, Estrin D (2001) Geography-informed energy conservation for ad hoc routing. In: Proceedings of the 7th annual ACM/IEEE international Conference on Mobile computing and networking (MobiCom01), Rome, Italy, July 2001
19. Karp B, Kung HT (2000) GPSR: greedy perimeter stateless routing for wireless networks. In: ACM/IEEE international conference on mobile computing and networking, pp 243–254
20. Yu Y, Estrin D, Govindan R (2001) Geographical and energy-aware routing: a recursive data dissemination protocol for wireless sensor networks. UCLA Computer Science Department Technical Report, UCLA-CSD TR-01-0023
21. Chang J-H, Tassiulas L (2000) Maximum lifetime routing in wireless sensor networks. In: Proceedings of the advanced telecommunications and information distribution research program (ATIRP2000), College Park, MD, March 2000

22. Kalpakis K, Dasgupta K, Namjoshi P (2002) Maximum lifetime data gathering and aggregation in wireless sensor networks. In: Proceedings of IEEE international conference on networking (NETWORKS 02), Atlanta, GA, August 2002
23. Ye F et al (2001) A scalable solution to minimum cost forwarding in large scale sensor networks. In: Proceedings of international conference on computer communications and networks (ICCCN), Dallas, TX, October 2001
24. Sohrabi K et al (2000) Protocols for self-organization of a wireless sensor network. *IEEE Pers Commun* 7(5):16–27
25. Akkaya K, Younis M (2003) An energy-aware QoS routing protocol for wireless sensor networks. In: Proceedings of the IEEE workshop on mobile and wireless networks (MWN 2003), Providence, RI, May 2003

## ***Further Readings***

26. Wireless Sensor Network. <http://en.wikipedia.org/wiki/WirelessSensorNetworks>
27. Estrin D, Govindan R, Heidemann J, Kumar S (1999) Next century challenges: a scalable coordination in sensor networks. In: Proceedings of the 5th annual ACM/IEEE international conference on mobile computing and networking (MobiCom'99), USC/Information Sciences Institute, pp 263–270, August 1999
28. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. *Comput Netw* 38:393–422
29. Al-Karaki JN, Kamal AE, (2004) Routing techniques in wireless sensor networks: a survey. In: IEEE wireless communications, Dept. of Electrical and Computer Engineering Iowa State University, Ames, Iowa, December 2004. <https://doi.org/10.1109/MWC.2004.1368893>
30. Akkaya K, Younis M (2005) A survey on routing protocols for wireless sensor networks. *Ad Hoc Networks*, 3(3), Elsevier B.V, pp 325–349, May 2005
31. Heinzelman W, Kulik J, Balakrishnan H (1999) Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings of the 5th annual ACM/IEEE international conference on mobile computing and networking (MobiCom'99), Seattle, WA, August 1999

# Drive into Future World Using Artificial Intelligence with Its Application in Sensor-Based Car Without Driver



Ridhima Sehgal

**Abstract** In the cutting edge period, artificial intelligence (AI) revolutionizes the world. Intelligent machines will alter the human capabilities in several areas. Artificial knowledge is the insight shown by different machines and software. It is considered as the subordinate of software engineering. Computerized analysis is turning into a renowned field in as it has upgraded the life of humans in many areas. Artificial insight over the last two decades has significantly improved the execution of the manufacturing and management systems. Study in the sector of artificial intelligence has offered ascend to the master framework, and its application in the vehicles is engaged to be computerized to give human driver loosened up driving. In the field of car, different viewpoints have been viewed as which makes a vehicle computerized. This paper discusses the artificial intelligence, types, space of artificial intelligence and its applications in field of the car without the driver using sensor-based technology and also talks about difficulty in the pathway of sensor-based cars and its future scope.

**Keywords** Artificial intelligence (AI) · Machines · Shrewd · Robots · Knowledgebase · Natural language processing (NLP) · Fuzzy logic (FL) · Human · Sensors

## 1 Introduction

AI describes the potential of machine learning just like human beings and the capability to respond to definite behaviours also recognized as artificial intelligence, or in other words, AI is the “replication of human intellect processes by machines, mainly computer systems” [1].

AI is a part of software engineering that intends to make smart machines. It has turned into a fundamental piece of the innovation business. As the world is advancing, researchers and specialists are attempting to take human life in a safer place. Smart

---

R. Sehgal (✉)  
Computer Science, BBK DAV College, Amritsar, India  
e-mail: [Ridhimasehgal2333@gmail.com](mailto:Ridhimasehgal2333@gmail.com)

machines will replace capacities of human beings in many regions. So, AI is the study and developments of smart machines and software which can reason, learn, collect information, correspond, manipulate and recognize the objects [2].

AI changed our typical lives. We are encompassed by this innovation from programmed stopping frameworks, intelligent sensors for taking marvellous photographs and individual help.

With Artificial Intelligence, you don't have to prearrange a machine to do some work, in spite of that you can make a machine with modified calculations which can work with possessing the knowledge, and that is the marvelousness of artificial intelligence.

## 2 Categorization of AI

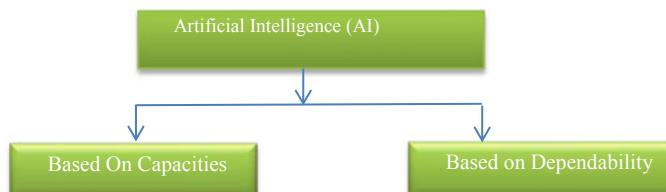
AI can be partitioned in different sorts. There are for the most part two kinds of initial order which depend on capacities and dependent.

Following is stream graph which clarifies the sorts of AI (Fig. 1).

### 2.1 Based on Capacities

Based on capacities, the AI can be classified into following categories:

- **Weak or Narrow AI:** Tight AI is a sort of AI which can play out a devoted undertaking with intelligence. The most normal and common accessible AI is Narrow AI. Slender AI cannot perform past its field or constraints, as it is prepared for one explicit errand. Henceforth, it is likewise named as feeble AI. Thin AI can flop in flighty ways in the event that it goes past its points of confinement. Some of the examples of Narrow AI include purchasing suggestions on e-commerce platforms, driverless cars, recognition of speech and recognition of images [3].
- **General Artificial Intelligence:** General AI is a kind of knowledge that could play out any human. The thought behind the general AI is to make such a framework which could be more brilliant and take on a similar mindset as a human itself.



**Fig. 1** Categories of artificial intelligence. *Source Author*

Presently, there is no such framework exists which could go under general AI and can play out any errand as immaculate as a human. The overall specialists are currently centred on creating machines with general AI. The frameworks with general AI are currently under research, and it will set aside loads of endeavours and effort to grow such frameworks.

- **Strong Artificial Intelligence:** Strong AI is a degree of astuteness of systems at which machines could outperform human knowledge [4]. It is a result of general AI. Some essential qualities of robust AI incorporate the capacity to think, to solve the riddle, make decisions, plan, learn and convey information itself. Super or strong AI is as yet a speculative idea of AI.

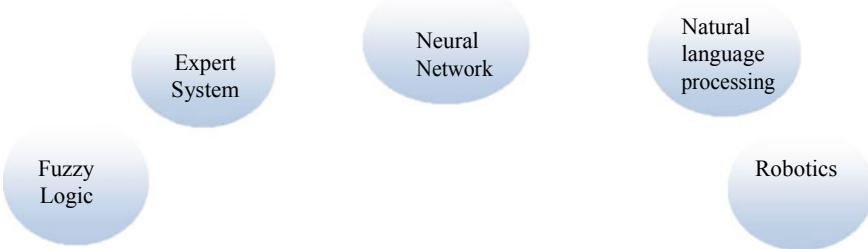
## 2.2 *Based on Dependability: Based on Dependability, the AI Can Be Divided into Following Categories*

- **Reactive Machines:** These machines spotlight on current situations and respond on it according to conceivable best activity, Google's Alpha Go is an example of reactive machines [5].
- **Limited Memory:** This type of artificial intelligence can store little information for a brief time frame. These machines can utilize put away information temporarily period as it were. Self-driving autos are probably the best case of limited memory frameworks. These vehicles can store an ongoing rate of close-by automobiles, the separation of different vehicles, speed limit and other data to explore the street.
- **Theory of Mind:** AI ought to comprehend human feelings, individuals, convictions, and have the option to connect socially like people. This kind of AI machines are as yet not grown; however, specialists are attempting bunches of endeavours and enhancement for budding such AI machines.
- **Self-Awareness:** It is the fate of AI. These machines will be hyper-savvy and will have their very own cognizance, estimations and mindfulness. These machines will be more brilliant than the human personality. Such AI does not actually exist and it is a theoretical idea.

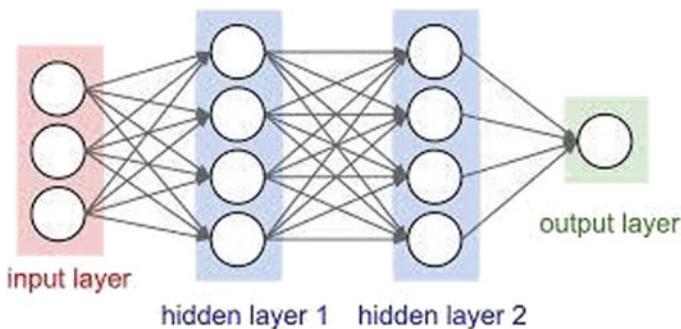
## 3 Spaces of Artificial Intelligence

Artificial Intelligence spreads plenty of domains. The various areas artificial intelligence covers are as following (Fig. 2):

- **Neural Network(NN):** A NN or artificial neural network is composed of artificial neurons or node connections of the biological neuron, and they are modelled as weights. A positive weight represents an excitatory connection, whereas a negative value represents an inhibitory connection. All inputs are modified by weights and added [6] (Fig. 3).

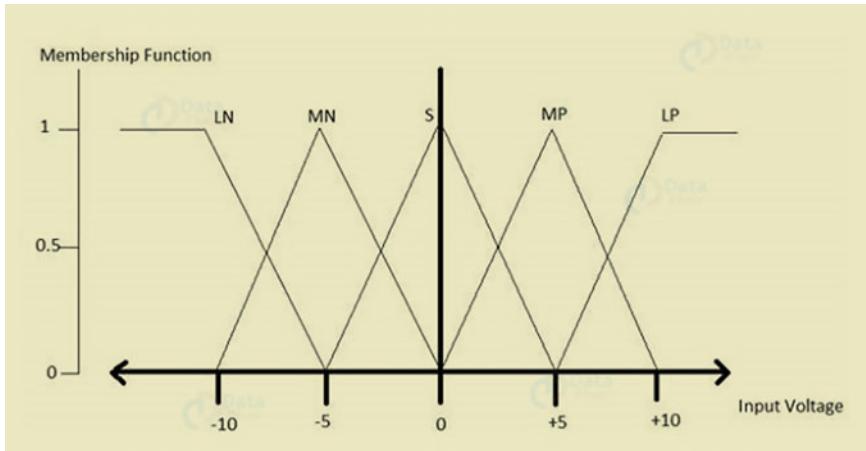


**Fig. 2** Spaces of artificial intelligence. *Source Author*



**Fig. 3** Neural network architecture. <https://cs231n.github.io/neural-networks-1/>

- **Expert Systems:** In AI, an expert system emulates the decision-making capability of humans [7]. It is a computer program that utilizes AI techniques to simulate the behaviour of a human having knowledge and experience in a meticulous field. It basically involves two parts, i.e. knowledge base and expert system shell. The Knowledge base is a gathering of guidelines encoded as metadata in a document framework, or all the more frequently in a social database. The expert system shell is an issue autonomous part lodging offices for making, altering and executing rules.
- **Robotics:** It is a space in AI that manages the exploration of creating smart and proficient robots. Robots are counterfeit specialists acting in a real domain. The Robot is planned for controlling the items by seeing, picking, moving and devastating it, thereby performing continuous tasks without getting bored or fatigued. Artificial intelligence robots are used in various fields like medical care, space exploration, movie industry etc.
- **Natural Language Processing:** Natural language processing (NLP) is the capacity of a system program to comprehend human language. NLP is a segment of computerized reasoning (AI). By making use of NLP and its components, one can systematize the vast chunks of text data, perform several automated tasks and resolve a wide range of troubles



**Fig. 4** Fuzzy logic system. <https://data-flair.training/blogs/fuzzy-logic-systems/>

such as machine translation, analysis of sentiments, recognition of speech and topic segmentation etc. The NLP involves making computers to execute essential jobs with the languages that humans use [8].

- **Fuzzy Logic System (FL):** The approach of FL imitates the way of decision-making in humans which involves all intermediary possibilities between digital values HIGH and LOW [9]. The usual logic block that a computer can understand takes specific input and produces a definite output as TRUE or FALSE. FL systems are used in a variety of applications like height control of shuttle, satellite height control, controlling traffic and many more (Fig. 4).

## 4 Artificial Intelligence Applications: Car Without Driver

Individuals around the globe are currently very much eager about the dispatch of self-ruling vehicles. The claim to fame of this vehicle is its capacity to see its condition utilizing the cutting edge type of AI, and make choices without the help of any driver [10].

These automobiles are outfitted with unique sensors, processors and another database which is in charge of the activity of this vehicle and does not require any driver. It explores itself following up to the goal point mentioned by clients. For sure, it is the enormous upset in the field of mechanical autonomy, which is contributing a great deal to make this planet a more secure spot.

On a specialized premise, this vehicle is structured dependent on the different zones of the building which incorporates electrical, mechanical sciences and control designing and so forth.

- **Easy Availability:** It will be never again when driverless vehicle would keep running over the street when AI, machine learning, deep learning and NNs innovation become well known. We tap on our cell phone application, and driverless vehicle is left on your entryway. By utilizing voice order, you will tell the vehicle for desired goal. The vehicle will utilize GPS, Camera and Google guide and Navigate itself over the street corner and traffic signal.
- **Lesser Road Accidents:** According to the survey, many people killed in road accidents which led to the development of car without a driver. This number can be enormously decreased by putting self-governing vehicles energetically, which are undeniably more dependable and respond quickly than people. It will likewise cause a decrease in the rush hour gridlock blockage, as the proficiency of self-sufficient vehicle makes it robust in a method for keeping little holes among vehicles, and its remarkable administration of speed and time. Following the route track without considering some other diversion makes it friendlier than the regular vehicles worked by drivers.
- **Boon for Non-drivers:** The AI-based specific vehicles can play out the majority of its outing with no inhabitant, and it is increasingly obliging for the individuals who are not ready to drive or might be because of different variables unfit to adapt to driving. Despite the fact that designers are putting part more exertion to make it as precise as could be expected under the circumstances, this innovation does not give that exhibition with the goal that it very well may be confided in blinded to put it out and about. So, these AI-based technology cars used sensor-based technology.

One of the real worries of these autos is to receive the more secure system to perform overwhelming securely, so thus there are number of various sensors and stereo cameras where they recognize the vehicle ahead and the separation, connected it to the speed of the vehicle itself and by performing other discernment through sensors it completes the situation to pass the other vehicle.

## 5 Hurdles in Pathway of a Car Without Driver

No doubt, it is clear the different favourable circumstances of this driverless vehicle like giving the wellspring of versatility for non-drivers and diminished the driving worry for the driver, however alongside these valuable impacts, there are quantities of difficulties and troubles that are associated with the execution of this innovation [11].

Following are a portion of the pressing issues related to the driverless vehicles:

- **High Price:** High prices of such cars are a big challenge of using it. As the cost of such cars is very high so therefore all persons cannot afford them.
- **Risk of Intruders:** Security and protection are continually being the most significant issue related to the electronic framework. Self-ruling vehicles depend on the

AI framework, where it additionally requires a wellspring of the internet for overseeing and data trade, and this is the trade-off medium which can be mishandled by the programmers.

- **Jobless Drivers:** With using sensors-based high technological cars, one of the pressing issues is the joblessness for the drivers. The execution will cause the supplanting of all the manual strategies of driving which explicitly incorporates for cabs, trucking and so on, and it is the wellspring of extreme work for many people groups the world. Due to this, drivers around the world will become jobless.
- **Failure of Sensors:** Autonomous vehicles depend vigorously on their sensors and sometimes these sensors may become inoperative. These sensors may become inoperative under following conditions:

Radical climate conditions.

Barriers and neighbourhood transit regulations.

- **System Crash:** Like all system frameworks, self-ruling autos would be modified to run a specific way; however, with every single robotized framework, there is consistently a danger of hacking or slamming. No framework is faultless, and if a programmer had the option to get into the vehicle's product, they could reconstruct the vehicle to do any number of things.
- **Driverless Vehicles Cannot Decipher Human Traffic Signals with Ebb and Flow Innovations:** Our momentum utilization of driverless vehicles works utilizing an arrangement of cameras, radar and lidar sensors. This innovation makes it feasible for the PCs of the vehicle to "see" the earth around them, distinguish traffic, or stop when it experiences an impediment. There are times when crisis circumstances require law implementation, utility labourers, firemen or other specialists on call for direct traffic-utilizing hand signals. In the event that a driverless vehicle was to experience such a circumstance, at that point, it would not realize what to do.

The various hurdles can be removed by developing efficient sensors which can work under extreme conditions too. However, these driverless cars cannot be implemented fully because of traffic conditions so there will always be a requirement of humans.

## 6 Conclusion and Future Scope

Though it has pros and cons of sensors-based car without a driver but, we cannot hide the fact that it replaces the conventional manual driving system. Regardless, as processing power, detecting limit, and remote system for vehicles rapidly increase, helped driving and proactive security forewarning is speeding towards this present reality. Thus, it helps to minimize road mishaps which occur due to human negligence or mistake. It is significant that car organizations put all their push to make it as secure as conceivable in light of the fact that any mishap brought about by these vehicles

can upset all the business. Be that as it may, in reality, this innovation can have a vital impact as the wellspring of transportation for open and military, in the different quest activity and use for getting to a specific area which is somewhat a hazard for the human driver. This advanced technology would take some time to adopt it while keeping security threats and concerns in mind.

## References

1. Pabby G, Kumar N (2017) A review on artificial intelligence, challenges involved & its applications. *Int J Adv Res Comput Eng Technol (IJARCET)* 6(10)
2. Pannu A (2015) Artificial Intelligence and its applications in different areas. *Int J Innov Technol (IJEIT)* 4(10)
3. <https://www.javatpoint.com/types-of-artificial-intelligence>. LNCS Homepage, <http://www.springer.com/lncs>. Last accessed 21 Nov 2016
4. McDermott D (2007) Artificial intelligence and consciousness. *The Cambridge handbook of consciousness*, pp 117–150
5. Lu H, Li C, Ki M, Serikaw S (2018) Brain intelligence: go beyond artificial intelligence. *Mobile Netw Appl* 23(2):368–375
6. Kumar N, Gupta S (2016) Offline handwritten Gurmukhi Character recognition: a review. *Int J Softw Eng Appl* 10(5):77–86
7. Shafer G (1987) Probability judgment in artificial intelligence and expert systems. *Stat Sci* 2(1):3–16
8. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. *J Mach Learn Res* 12(8):2493–2537
9. Xiang X, Yu C, Lapierre L, Zhang J, Zhang Q (2018) Survey on fuzzy-logic-based guidance and control of marine surface vehicles and underwater vehicles. *Int J Fuzzy Syst* 20(2):572–586
10. Lugano G (2017) Virtual assistants and self-driving cars. In: IEEE international Conference on ITS telecommunications, pp 1–5
11. Murray R (2007) Driverless cars. *Comput Control Eng* 18(3):14–17

# Linking and Digital Story Telling Approach in Teaching Towards Enhancing and Engagement of Smart Study



Sanjay Tejasvee and Manoj Kuri

**Abstract** Today, there are many new technological devices such as smart board, projector, computer, camera, scanner and much useful software are available to teach the students in efficient way. By using these tools and devices, educators get more opportunities to enhance and elaborate their knowledge; standard and skill with students and students also feel comfortable and well engaged, motivate and achieve good outcomes. Linking and digital storytelling aspect perfectly worked to enhance and makes interesting teaching methodology with full engagement of learner while class is going on. Education system still suffers from many challenges such as to enhance student's engagement for better results, how to use innovative instructional approach to engage students. Aim of this research paper is to introduce and present constructive part of linking and storytelling approach. This paper will also explore the standardization, efficiency, impact and challenges in executing linking and storytelling approach. Linking and digital storytelling method is powerful technique to integrate learning process with knowledge activities to produce more attractive and moving commanding educational surrounding. Thus, move towards in this way will be potential to improve learner's engagement and deliver enhanced education-related results. This paper is also trying to present the inclination in education in modern technique that is smart learning prevailing in education by the execution of smart learning.

**Keywords** Technological · Digital · Technologies · Storytelling · Methodology · Enhance · Integrate and engagement

---

S. Tejasvee (✉)

MCA Department, Government Engineering College Bikaner, Bikaner, Rajasthan 334004, India  
e-mail: [drsanjaytejasvee@gmail.com](mailto:drsanjaytejasvee@gmail.com)

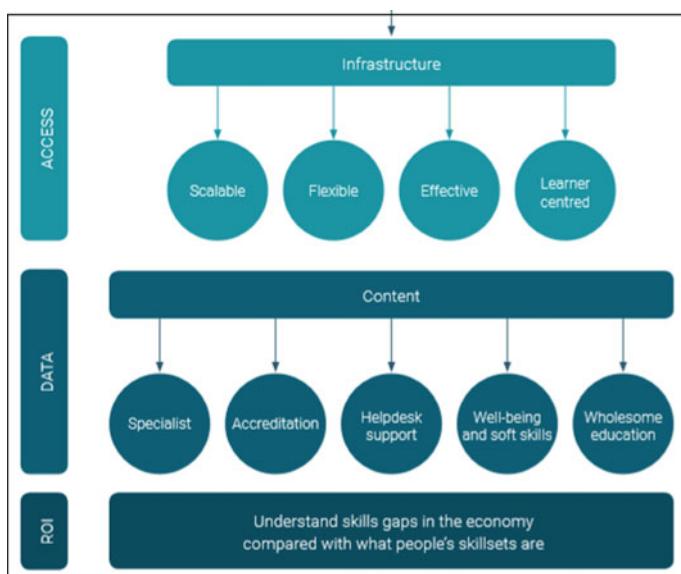
M. Kuri

Department of Electronics and Communication, Government Engineering College Bikaner,  
Bikaner, Rajasthan 334004, India  
e-mail: [Kuri.manoj@gmail.com](mailto:Kuri.manoj@gmail.com)

## 1 Introduction

Linking and storytelling can not only be used as a technique to teach earlier traditions and cultures but also can be used in better professional and academic study. Linking and storytelling approach also operate as a tool to convey skills in as social context. Today new technologies are gradually using to deliver education and skills in new innovative manner. Today, latest digital resources are using and rapidly increase to practice in education system. By the utilization of high speed changing digital tools and technologies in the education system, new skills will be emerged.

Latest technologies contribute to transforming learning and skills development and as the changes of life style in modern society, it is needed to be required some changes in the educational techniques as compared to traditional methods of teaching. The technological advancement is one of the most important factors to enhance the educational system with engagement of student while studying. Technological advancement means share the skill by the use of several tools and devices such as digital cameras, software and authoring tools to help students in constructing their own knowledge and ideas to present and share them more effective [1]. According to Armstrong, digital cameras, computers, editing software and other technological tools are becoming more eagerly reachable in the classrooms and present learners and teachers with the tools to create digital stories more simply than ever before [2]. Smart class also allows teachers to right away assess and estimate the learning accomplish by their students in class with an inventive assessment technology and smart evaluation system. Smart class is powered by a vast depository of digital



**Fig. 1** Initial frameworks for inclusive education in the digital world [9]

instruction resources exactly mapped to meet the explicit objectives laid out by special state learning standard [3]. An innovative and constructivist approach indicates to the value of mentors as key drivers for the online collaborative learning. Information and communication technology powerfully assists the learning process by the utilization of images and audio which can create their story and connect in an in depth learning [4]. Digitalize storytelling is a one technological application that is fine situated to obtain benefit of user-guided content and to assist teachers to use technology efficiently in classes [5].

## 2 Constructive Aspects

In late 1980s, digital storytelling approach appeared as a technique employed by theatre workers of community for enabling recording, producing and disseminating etc. in California [6]. “A short story was only two-three minutes long, where the storyteller uses his own voice to tell his own story. The individual element is emphasized and can be associated and linked to the other persons, a place to interest or to something that will feel the story an individual touch” [7]. This has developed in a number of ways, shaped by advance in personal computing and recording technology and use in variety of educational and non-educational contexts [7, 8].

In today’s digital era, linking techniques using digital teaching resources plays a vital role in education system. Most of person captures the way of smart classes in these days. It is true that that if a matter is understood by visual and image methods, it becomes more helpful to recognize to learners. So the requirement of storytelling method using smart digital devices and ICT tools is being convincingly raised. “Smart Study” serves better education by videos and presentations. Many researchers already found that a student or learner can learn and train in better manner by visualization. At a time, all students may not recognize the teaching methodology of a educator, but can well identify and learn by linking and digital storytelling method. Yes, off course, it was seen in many cases that movies, visual clips and audio clips etc. are captured in students brain better than the topics taught in classroom as traditional methods. Linking study material with live example and storytelling technique with ICTs creates an high attention and interest in students and learners. So digital learning with linking and storytelling aspect is absolutely far better to help for sustaining the interest and engagement and focus of students completely for the intact learning experience. Some of the major constrictive points for linking and digital storytelling approach in the learning and education system are as follows:

- Effective learning as the information is strongly embedded in leaner’s brain by audio visual senses appeal.
- Long time storage in the learner’s brain can retrieve easily with help of linking techniques.
- No wastage of time in the classroom to draw diagrams and write large syntax etc., thus time is utilized more for the active and smart learning.

- Students go through a virtual visit or trips while covering a topic which is useful to develop interest and link objects sequential also.
- In digital teaching, educator uses special markers to mark an important spots while teaching to formulate and clear the concept.
- Digital teaching has smart board that has an inbuilt library which enables educators to have an instant look at it in case of requirement.
- Storytelling and digital teaching technique lead to active learning where both the educators and leaner are involved with their full focus.

### 3 Standardization and Impact

It should be identified that how teaching technique can help students and learners to enlarge their engagement while studying and develop interest in them. It is the main objective of any educator to teach and educate their students with full knowledge and complete clarification of the concepts whatever they deliver.

There are very limited options rather than linking and digital storytelling method to improve consistent record and inform formation agenda help to create their atmosphere [10]. Digital learning is an instruction carried out that efficiently uses technology to make stronger a leaner's skill experience. Digital learning covers a wide range of tools and carry out, including, between others, online and decisive consideration. In exacting, blended education occurs any time a student learns, at least in fraction, at a supervised element and mortar place away from home and at least in part by online deliverance with some element of student control over time, place, path [11]. Emerging substantiation designate that cognitive, intrapersonal and interpersonal competence can be educated and academic in ways that support transfer. Digital education put emphasis on not only comfortable actuality but also how, when and why to apply this knowledge is necessary to transmit [12]. Education point of view linking and storytelling method use personification of ICT resources. This method is related to our lives or live examples so this method is an significant part of teaching as well as leaning. It has been seen in the few past years some educational institutes have already been walking around through the approach of linking and storytelling by digital medium. The adoption of this teaching and learning approach not only expresses learning in expression of results but also brings their significant executions for all portions of curriculum and lecture plan design, deliverance, representation, evaluation and review standard. Uses of linking and storytelling approach in education can increase various skills of learners and can enhance many learning skills including writing, designing, communicating and research. In addition, this approach can help to the learners with works they earlier found critical including formation of sentences and text body. This approach overcomes their writing problem also.

Even though tools of ICTs is generating many opportunities for basic changes in the way of teacher's teaching methods with student's learning manner, it is observed that only one third of educators ready to apply it efficiently. Many software packages

like word, spreadsheet, presentation and internet browsers help the educators to increase their efficiency by making reports, lecture plans, notes and communicate with their associates and guardians of the students [13, 14]. Storytelling and linking technique are not a novel approach, and numerous people already identified it as a powerful impact on building strong framework in the educational field.

This technique is used as boost for the innovations for live development which is also one of the main objectives of education. This technique provides the micro learning, and human brain catches efficiently when knowledge is delivered in smaller fragments. This mechanism also provides better gain insights of practical understanding in their relevant ground. In this approach, digital textbooks and e-books used which provide a faster text searching and navigation in relevance categories and transitioning to the format of digital are cost effective and easier for leaner and educator both. Thus, it can be said that this approach has effectiveness, interoperability, interactivity and flexibility in the learning, teaching and education sector.

## 4 Challenges in Execution

In the area of formal education, learners focused on the significance of making the difference between using technology to learn rather than teaching how to utilize technologies. It should be noted that learning could be achieved in several ways, the importance of technology is separate part which can support powerfully in teaching and learning, it cannot replace the teacher [15–17]. However, many educational institutes hire only those people who are able to teach in smart manner with linking and storytelling mechanism and can also efficiently use technological tools to teach students and today many educators use this phenomenon and work towards put up this mechanism in trend. Several resources and tools are available for powerful teaching and enhance teaching and learning procedures. Still, there are some challenges and issues come in front of execution of linking and storytelling such as: [18–20]

- Linking and digital storytelling can be a dominant tool for opportunity to carry out within effort organization, but converse any kind of message into appropriate link and create a immediate storytelling is critical.
- Educators have to do put on extra efforts at initial level due to first they have to develop the linking procedure and have to innovate how to create interesting story of a study topic? Sometimes, it is very critical to create link to real world and make a rational story.
- Inadequate and less availability of educator's training to educate students which enhance the educator's logic to produce link and make story on subjective matter.
- Lack of proper monitoring is also a reason which create problem in execution, maintain and observation of learners.
- Explain, show linking, telling story in detail then analysis of all these things take much time at various level.

- Some traditional educators and educational organization do not appreciate educator who are using linking and storytelling technique for students or learner. This thing demotivates the educator to use this technique. So, improper appraisal is also impacted in adverse.

## 5 Conclusion

Linking and storytelling approach is not a novel concept or form in the education field. People always told the stories and tried to link with many aspects while passing their tradition and custom to the new generation. Linking and storytelling technique provide flexibility and interactivity which create a connection between educator and learner. This approach engage the learner and educator feel a different kind of experience of literacy every time whenever educator teaching with this approach. Linking and storytelling approach develop the better imagination and visualization power of learners. This approach provides the better students as well as the better teacher due to learning through interesting manner. The link and storytelling method are almost certainly the most fundamental memory technique and are extremely easy to understand and use. This technique explains the matter and material of study by coding information into images then linking these images together vice versa. This assists to keep events in a logical order and can improve learner's ability to identify and memorize things which delivered wherever anybody forget then sequence of images and linking help to retrieval.

## References

1. Standley M (2003) Digital storytelling: using new technology and the power of stories to help our students learn and teach cable in the classroom 16–18
2. Armstrong S (2003) The power of storytelling in education, in snapshots. In: Armstrong S (ed) Snapshots educational insights from the thorn burg centre, pp 11–20
3. Sanjeev K (Trained Graduate Teacher in Non medical), E-Learning and role of smart class rooms in education in new era of technology
4. Pounsford J (2007) Using storytelling, conversation and coaching to engage: how to initiate meaningful conversations inside your organization. Strat Commun Manage 11(3):32–35
5. Robin BR (2008) Digital storytelling: a powerful technology tool for the 21st century classroom. Theory Pract 47(3):220–228
6. Lambert J (2009) Where it all started: the center of digital storytelling in California, story circle; digital storytelling around the world, pp 79–90
7. Normann A (2011) Digital storytelling in second language learning (Noreweigan, University of science & technology, Faculty of Social Sciences and Technology Management 2011), p 125
8. Clarke R, Adam A (2012) Digital storytelling in Australia academic perspectives and reflections. Arts Humanit High Educ 11(1–2):157–176
9. Report; Digital learning; Education and skills in the digital age; sarah Grand-Clement; An overview of the consultation on digital learning held as part of the corsham institute. Through leadership programme 2017

10. Dusan K (2010) E-learning, Sarajevo
11. Digital Learning Day website. <http://www.digitallearningday.org/>
12. National Research Council. Education for life and work developing transferable knowledge and skills in the 21st Century, July 2012
13. Pellegrino JW, Chudowsky N, Glaser R (2001) Knowing what students know: the science and design of educational assessment. Washington, DC: National Academy Press
14. Reil M (2000) New designs for connected teaching and learning. White paper commissioned for The Secretary's Conference on Educational Technology Evaluating the Effectiveness of Technology, Washington, DC, 11–12 Sept 2000
15. OECD (Organization for Economic Co-operation and Development), Students, computers and learning: making the connection. Programme for International Student Assessment report. Paris: OECD Publishing. <https://doi.org/10.1787/9789264239555-en> (2015)
16. Technical Report: results from the study: student 'use' of 'digital' learning 'materials' implications for the NSDL <https://pdfs.semanticscholar.org/3473/0d824aa708ca7d5398b4907b5353e2fc41fc.pdf>
17. VanderArk T, Carri S How digital learning contributes to deeper learning. <https://www.gettingsmart.com/wp-content/uploads/2012/12/Digital-Learning-Deeper-Learning-Full-White-Paper.pdf>
18. Smeda N, Dakich E, Sharda N (2014) The effectiveness of digital storytelling in the classrooms: a comprehensive study; smart learning Environments
19. Nunes R, Cruz J, Miguel Martins, Sousa MJ (2017) Digital learning methodologies and tools—a literature review" conference paper. <https://www.researchgate.net/publication/318679851>
20. Ashfaque et al (2014) Trends in education smart learning approach. Int J Adv Res Comput Sci Softw Eng 4(10):319–327

# Classifying Titanic Passenger Data and Prediction of Survival from Disaster



Shashank Shekhar, Deepak Arora, and Puneet Sharma

**Abstract** The sinking of the ship named Titanic is one of the most historic shipwrecks in the world. It was held on April 14, 1912. Thousands of people died in this accident. Out of 3000 passenger, almost 1500 cause death in this accident. The reason behind this accident is due to less lifeboat because they never thought that this ship would ever sink because it is one of the largest ships in history at that time. So in this paper, an analytical approach has been proposed by the authors in order to predict the survival rate of people on the Titanic ship. For the experimental study, authors have selected Titanic dataset and applied suitable classifiers with the help of Python programming. For study purpose, spot check algorithm has been applied to predict what kind of people was survived. The experimental results have shown the model prediction value around 86.29% through spot check algorithm which found most satisfactory over results found in the literature varied from 72 to 82% only.

**Keywords** Logistic regression · Data mining · Sport check algorithm · Statistics · Probability · Linear regression

## 1 Introduction

The most famous incident happened on April 12 is the Titanic shipwreck accident, which causes thousand of passenger death because of this shipwreck accident. The reason behind this accident is because of collision with iceberg. They thought that this ship would never sink so they kept very less lifeboat, and because of less lifeboat,

---

S. Shekhar (✉) · D. Arora · P. Sharma

Department of Information Technology, Amity School of Engineering and Technology, Amity University, Lucknow Campus, Lucknow, India

e-mail: [shekharshashank106@gmail.com](mailto:shekharshashank106@gmail.com)

D. Arora

e-mail: [deepakarorainbox@gmail.com](mailto:deepakarorainbox@gmail.com)

P. Sharma

e-mail: [puneetgrandmaster@gmail.com](mailto:puneetgrandmaster@gmail.com)

many people cause death. In Titanic ship there are many classes divided like first class, second class, and third class and fair in Titanic ship belongs from class only like people belongs to first-class have highest fair as compared to second class so the chance of survival of first-class have more and they have more lifeboat. By calculating the predicted value of survival, authors have applied the concept of machine learning and calculated the predicted value of how many people survive in that accident. Authors have taken both the training and testing data and from the dataset. In this research, work authors have used the selected feature that has been used to predict the survival rate of people and the rest of the instance and feature has been omitted. That feature and instances which having missing values, firstly to be filled with assumption taken from that data similarly in the case of age group, the missing value can be assumed by extracting the sir name like (Mr, Mrs, Dr, master, etc.). It can be extracted from the “name” column of historic data and title of the name. The aim of applying the concept of machine learning is to predict the accurate value of survival of passengers and the specific type of passengers. Authors have used the Python programming language for the analytical study. The classification accuracy will be checked by spot check algorithm, there are 10 types of spot check algorithms, author has applied all 10 spot check algorithms, and the classifier model showing best prediction rate will be taken into consideration.

## 2 Literature Survey

In the literature survey, authors have found that most of the research work is based on the prediction of survival of people and category which they belong to an adult, young age, baby or child, etc. The author has also seen that the previous research paper [1] and [2] predicted survival with random forest classifier. He predicted 80% survival rate. He has applied the concept of machine earning with different classifiers like random forest using R. Majorly, he has taken dataset from Kaggle and used five feature instances during his experimental study. Another research work proposed in [3] and [4] also indicates the usage of the random forest model to predict the survival of passengers and type of passengers. His best-predicted value is 79%. As per the literature, similar kind of experimentation can be seen toward all models of random forest classifier, and with the best-predicted model, they trained the machine. In this research work, authors have used the spot check algorithm to find the best-predicted value, and out of ten spot check algorithms, best-predicted algorithm model has been taken. The experimental results show the predicted value of survival of people is 86% which seems better-predicted value from others as compared to research work as mentioned in the literature. In the proposed work, authors have used Python programming for implementation purpose. This research work is focused on survival prediction of people through spot check algorithm, and experimental results show satisfactory results if compared to the existing literature work.

### 3 Datasets and Experimental Setup for the Study

From the Kaggle data repository, authors have taken two types of data. One is testing data, and the other is training data. The training data is used to train the prediction model, and the testing data is used to make our model find the predicted value and the best-predicted model. In this research work, authors have evaluated the proposed model, and after evaluating, the predicted results have been compared with others. The detailed description of the Titanic dataset and missing values for the entire dataset are given in Tables 1 and 2.

**Table 1** Dataset attribute description

Attributes	Description	Factors
Sex	M/F	Male or Female
Survival	Passenger survived	1 = YES 2 = NO
Passenger ID	Unique ID	1, 2, 3, 4, 5, 6, 7...
Ticket	Ticket PNR number	
Passenger Class	Defined class	Cabin Bool-1 Cabin Bool-2 Cabin Bool-3
Age	Age of passenger in(years)	Extracted from their Sir name
Embarked	Port from where passenger started their journey Q for Queen Town C for Cherbourg S for Southampton	Q, S, C
Parch	All parents and children aboard Titanic	
Fare	All passenger fare	

**Table 2** Missing values in the Titanic dataset

Attribute	No. of Missing Values
Name	0
Sex	0
Age	177
Sibsp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

PassengerId	Survived	Pclass	Sex	SibSp	Parch	Embarked	AgeGroup	Title	FareBand
1	0	3	0	1	0	1	4	1.0	1
2	1	1	1	1	0	2	6	3.0	4
3	1	3	1	0	0	1	5	2.0	2
4	1	1	1	1	0	1	5	3.0	4
5	0	3	0	0	0	1	5	1.0	2
6	0	3	0	0	0	3	5	1.0	2
7	0	1	0	0	0	1	6	1.0	4
8	0	3	0	3	1	1	1	4.0	3
9	1	3	1	0	2	1	5	3.0	2
10	1	2	1	1	0	2	3	3.0	3
11	1	3	1	1	1	1	1	2.0	3
12	1	1	1	0	0	1	6	2.0	3
13	0	3	0	0	0	1	4	1.0	2
14	0	3	0	1	5	1	6	1.0	4
15	0	3	1	0	0	1	3	2.0	1
16	1	2	1	0	0	1	6	3.0	3
17	0	3	0	4	1	3	1	4.0	3
18	1	2	0	0	0	1	5	1.0	2
19	0	3	1	1	0	1	5	3.0	3
20	1	3	1	0	0	2	6	3.0	1
21	0	2	0	0	0	1	5	1.0	3
22	1	2	0	0	0	1	5	1.0	2
23	1	3	1	0	0	3	3	2.0	2
24	1	1	0	0	0	1	5	1.0	4
25	0	3	1	3	1	1	2	2.0	3
26	1	3	1	1	5	1	6	3.0	4
27	0	3	0	0	0	2	5	1.0	1
28	0	1	0	3	2	1	4	1.0	4
29	1	3	1	0	0	3	4	2.0	1
30	0	3	0	0	0	1	5	1.0	1
..	...	...	...	...	...	...	...	...	...

**Fig. 1** Experimental data preparation

From the above data, authors have observed that in age column there is 177 missing value; in-cabin column there is 687 missing value; and in the embarked column there is 2 missing value, and the age feature is missing almost 19.8%; for finding the predicted value, we require age column completely. So in this column, we first fill the value through predicted value. By calculating from historic data, the author came to know that female has more chance of survival than men because the female survival rate is 74.2% and men survival rate is 18.8%; we can easily know that the survival rate of the female is more (Fig. 1).

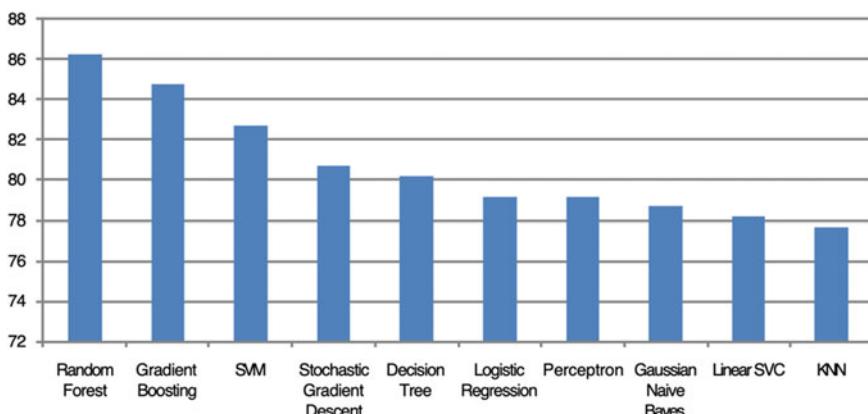
In age feature, there are 177 missing value almost 19% of data is missing; for prediction, age group is important, so for this we extract the Sir name from the passenger name like (MR, MRS, DR, MAJOR, SIR, etc.); from that, we predict age value according to the Sir name and fill all the missing value. We also calculate the survival of passenger according to cabin allotted to passenger so passenger and author came to know that passenger sitting in Cabin Bool 1 has survival rate 66.6% and passenger sitting in Cabin Bool 2 has survival rate 29.9%, so author can easily predict that in-cabin Bool 1 has more facility and they have very high rate of survival and this survival the author can also predict from passenger ticket fare the passenger fare which has high rate can belong to cabin Bool 1 as compared to passenger sitting in Cabin Bool 2.

After putting all the values in all feature and instances, we take 10 features and drop rest of features because we did not require those features in predicting the value. We take all values in numeric form, and now we calculate the predicted value by using spot check algorithm.

**Table 3** Classifier accuracy comparison

Classifier	Accuracy
Random Forest	86.29
Gradient Boosting Classifier (%)	84.77
Support Vector Machines (%)	82.74
Stochastic Gradient Descent (%)	80.71
Decision Tree (%)	80.20
Logistic Regression (%)	79.19
Perceptron (%)	79.19
Gaussian Naive Bayes (%)	78.68
Linear SVC (%)	78.17
KNN (%)	77.66

The author use 10 spot check Algorithm to test the data so that we get predicted value of survival of passenger and which type of passenger are they and the model classifier are Logistic Regression, Gaussian Naive Byes, Support vector Machine, Linear SVC, Perceptron, Decision Tree Classifier, Random Forest Classifier, KNN or K-nearest neighbors, Stochastic Gradient Descent, Gradient Boosting Classifier and for each model, the author has taken 80% of training data and check it on 20% of training data and find accuracy of survival of passenger. The accuracy performance comparison is given in Table 3 and Fig. 2 respectively.

**Fig. 2** Classifier accuracy comparison

## 4 Results and Discussion

After using all 10 spot check Algorithm the author came to observe that Random forest Classifier is giving the best-predicted survival rate that is almost 86.29% which is much better from the other Research paper-like Kakde [5] and Applicationsor Yu Cheung [6] etc. have calculated their predicted value from 75 to 80% which is less from the author predicted value. They have also chosen the random forest classifier to predict their survival rate, but In WEKA dataset also, their predicted value is less form the author.

## 5 Conclusion and Future Scope

While Analyzing Titanic data the author came to know that Random forest Analysis is giving the best predicted value and also the Research paper like [7] and [8] published on WEKA or in Data mining set their prediction value is between (75–81%) but the author worked more on this data and my prediction value is increased almost 5% from before, i.e., 86.29%, So spot check algorithm is the best one to test the data. Further, this research work can help in this to have a growth of almost 5% by using the best dataset to find the best-predicted value and the author can also observe that by taking different feature our prediction of survival of passenger may change. From this data, we also observed that survival of female is more than the survival of men, calculated with the concept of machine learning, and found a female percentage of survival is almost 75% and men percentage of survival is almost 18%. The sex feature is most essential in calculating the predicted value. From this data, we also came to observe that the survival of passenger also depends on the cabin Bool because when we calculate with the help of machine learning concept the author came to know that the passenger who is sitting in cabin Bool 1 has much higher survival rate than the passenger sitting in cabin Bool 2. The passenger sitting in cabin Bool 1 has almost 66% of predicted survival rate, and the person sitting in cabin bool 2 has survival rate almost 19%; so from this, we can say that Cabin Bool 1 has more lifeboats and they have given more facility as compared to another cabin bool and we can also say that cabin Bool 1 is safer as compared to any other cabin. This Titanic project can help to know about all the safe side in any ship, and the most important future scope is to help and make us learn all the concepts of machine learning from the starting or from very basic. In future, it helps also to find the more best-predicted value of survival of passenger by using different algorithms or we can also take the different features to find the another best feature in finding the best-predicted value. In future, we can also come up with the new concept or with the new idea so that we can come up with some best-predicted value by changing with attributes or by changing the value from its feature to get the best result. In the future, we can also come up with the new language which can be much easier from this way and can also be much simpler to implement in machine learning program data. Every time if

we can analyze the predicted result, so we can observe that day by day its predicted value is increasing so in coming future it is going to increase only and helps to get the best result.

## References

1. British Trade Administration (1911) British Wreck Commissioner's inquiry Report from <http://www.titanicinquiry.org/BOTInq/botRepBOT.php>, May (2012)
2. Lam E, Tang C (2012) Int J Comput Appl
3. Ali J, Khan R, Ahmad N, Maqsood I (2002) random forest and decision trees. Int J Comput Sci Issues (IJSCI), April 2012
4. Stevans L, Gleicher DL (2004) Int J Maritime History Dec 2004
5. Kakde Y (2018) Predicting survival on titanic by applying exploratory data analytics and machine learning techniques. Int J Comput Appl 179(44):32–38
6. Yu C, Cheung (2015) Titanic predicting the titanic survival rate, a random forest approach 173(32):21–26
7. Witley MA (2015) Survival of passengers on the RMS Titanic by Micheal Aaron Witley
8. Bhargava N, Sharma G (2013) Decision tree analysis on J48 algorithm for data mining, vol 3, June 2013

# Soft Skills: An Integral Part of Technical Education



Nisha Srivastava and Manoj Kuri

**Abstract** Soft skills also known as social or people skills is the interpersonal quality that a student should possess; it is a very important attribute for the engineers to get into the job. They have become the integral part of one's life in their workplace. It is something that one cannot learn by any training program or in the classroom. It can only be acquired through education, work, and life experiences. The present study deals with the soft skills like communication skills, positive attitude, teamwork, leadership quality, presentation skills, and interviews skills.

**Keywords** Engineers · Soft skills · Presentation skills · Communication skills · Positive attitude

## 1 Introduction

According to Cambridge dictionary, soft skills are “people’s abilities to communicate with each other and work well” [1]. Soft skills are skills related to behavior, expressed verbally or otherwise, as individual or in a group with other individual or a group in a given situation. They are soft as they have to be internalized not involving an external physical tool or device. With the help of this definition, we come to know that how important it is for the new candidates entering in the organization for the job, for which they need to develop confidence and awareness. An engineering student of the present era should focus on soft skills, communication skills, intrapersonal and interpersonal skills along with technical and hard skills because he has to prepare himself to attend placement selections, getting admissions into higher educational

---

N. Srivastava (✉) · M. Kuri  
Engineering College Bikaner, Bikaner 334004, India  
e-mail: [nshsrivastav@rediffmail.com](mailto:nshsrivastav@rediffmail.com)

M. Kuri  
e-mail: [kuri.manoj@gmail.com](mailto:kuri.manoj@gmail.com)

institutions for career advancement. He not only needs to train himself but also gets ready to face the challenges in the job; therefore, to begin with employability, it is very much an integral part of one's self and life. The student needs to be aware of his strengths and must develop the specific skills.

A student of engineering should understand the urgency behind improving communication skills and soft skills because after getting a job his focus will be shifted from specific to multifaceted, as in workplace he is required to do different tasks to prove his efficiency. Engineers and technocrat are no exceptions today; they along with their technical expertise are expected to have a broader general knowledge, awareness, and different skills for their success and the progress of the organization.

## 2 Importance

Soft skills like communication skills, leadership quality, presentation skills, team work, positive attitude, etc., are so very important for a graduate who is entering into the world of professionalism. A candidate must possess the quality of speaking fluently and confidentially as stated by a famous English poet Robert Frost: "Half of the world is filled with people who have something to say and cannot say it and the other half, of people who have nothing to say and keep on saying it" [2]; similarly, a student of engineering in today's world is also lacking with this quality as stated by Frost. In order to overcome these obstacles, the following points should be followed on the part of the candidate, with the help of which he/she can develop himself professionally and personally as well. Firstly, he/she should be responsible as responsibility is a confidence booster; secondly, he/she must have a check on the gestures and should make correct use of body language, and above all, one should always have a smiling face and a positive outlook. Same are the qualities that a leader should have while working as a team; a leader is a person who should, "seek first to understand then to understood" [3] as said by Covey.

## 3 Objectives

The main objective to learn the soft skills is to improve the personality of an individual in all spheres of life as according to Tony Robbins, "The way we communicate with others and with ourselves ultimately determines the quality of our lives." [3]. Soft skills are as important as are the hard skills, and like any other technical or hard skills, the need of it is equally important for the betterment and upliftment of an individual. It is becoming essential in today's world because there are variety of situations which everyone has to face. Today, people expect quick, smarter, and soft behavior in a stressful world. They make use of technology to communicate and present themselves, which requires more skills. Since it is a global world now, the

interactions are not limited; rather they are with the wider variety of culture and practices, and hence, training of such skills has become necessary.

There are several points that we can be understood as constitutions of soft skills in general terms.

1. Functions of Soft skills: Good manners and public behavior.
2. Greeting and introducing guest at any place.
3. Telephone etiquettes.
4. Appreciating and praising people.
5. Empathy and sympathy.
6. Leadership quality.
7. Presentation under different situations.
8. Grooming and general etiquette of eating and hygiene.
9. Time management and stress management.

With the help of these, a candidate can, not only raise his graph in work place but also groom his overall personality. The main points that are to be kept in mind in functional terms are as follows.

**Communication skills.** According to Brian Tracy “Communication is a skill that you can learn. It’s like riding a bicycle or typing. If you’re willing to work at it, you can rapidly improve the quality of every part of your life” [4]. Today, competency in English communication facilitates the students to analyze and summarize the work on a set time, to make all the possible arguments for a case, to work individually and in a group, and to think over the matters thoughtfully. Engineering graduates can bring all skills to their places of employment, together with sensitivity to language and literature. This makes them good candidates for a wide range of career including technical, personnel, administration, management, marketing, finance, and the media. The globalization in the business has taken the shape of a revolution in the field of academics and industry. Today, the world is known as a global market where everything is based on the outcome of the results, which has become a challenge for the students. Therefore, to achieve success in such a competitive world requires not only the knowledge of hard skills but also the talents in all types of skills.

**Presentation skills.** Presentation is a process of facing the public and putting up one’s own ideas and thoughts with the help of various aids like power points, newspaper cuttings, and visual aids. It consists either as a public speaker or representation of an institution, social behavior, and work-life balance. Presentation for some of the students is like a nightmare; it is a fear of phobia for most of the candidates because they do not know what presentation actually is, as said by Abraham Maslow, “If you plan on being anything less than you are capable of being, you will probably be unhappy all the days of your life” [5].

A candidate while giving the presentation should follow the following points:

Firstly, he/she should know the purpose of his presentation/communication, secondly what are the expectations of the audience, thirdly how to fulfill those expectations, and lastly one should be concrete, specific, practical in his/her ideas and objectives.

**Leadership.** Leadership is the process through which a leader influences the values, behavior, and attitude of others.

1. Leadership qualities can either be innate or can also be acquired.
2. A leader motivates the team members to perform well.
3. A leader also motivates the members in case of failure.

There are certain values that a good leader should possess.

1. Fairness, to treat people fairly and to give equal opportunities.
2. Freedom, to give freedom for expression of thoughts and ideas.
3. Commitment towards work.

Above all this, it becomes the responsibility of the leader to overcome the current situation; even if one is not in the management position, then also he must have the quality of a leader because leadership is such a quality where one must know the difference between what is right and what is not right. Mark Twain in one of his works said that “The difference between the almost right word and the right word is really a large matter—it’s the difference between the lightning bug and the lightning” [6].

## 4 Methodology

By following this specific methodology, while applying for the job, a student can make his entry easier in the technical workplace.

1. One should add skills to his curriculum vitae to specify the work done.
2. Always mark the skills in the covering letter.
3. While giving the job interview, tell about your skills with the help of various examples.

## 5 Review

Human perceptions in soft skills determine the proper understanding in relationships because perceptions are bound to be different, often contradictory; therefore, one should take proper care of his/her soft skills as one makes a point to have the command over his/her technical subject. It is beautifully stated by Oscar Wilde, “An optimist will tell you the glass is half-full; the pessimist, half-empty; and the engineer will tell you the glass is twice the size it needs to be” [6]. It makes no difference whether the glass is half filled or no water is there, but we should be happy that at least it is filled with something. Overall, we can say that “when you develop your emotional intelligence and live with high spiritual intelligence, your life will be so clear and transparent that even if someone speaks badly of you no one would believe it!” [6].

Danah Zohar in her Spiritual Capital wrote “To become better, deeper, more spiritually intelligent people, we have to grow a dimension of our being that is sensitive to the deepest meanings of human life—a sensitivity, if you like, to Plato’s famous triad of values: Goodness, Truth, and Beauty. We must live our lives as a vocation, as a calling to the service of those deepest values. To do that, we must act from the higher motivations that can drive human behavior. This is a long-term project, requiring tenacity and commitment” [7].

## 6 Conclusion

Globalization today is creating rapid increase in the demand for engineers who have the ability to deal effectively with professionals from different areas of the world. An engineering student of this modern era acquires so many skills that can be transferred to others in the field employment; his literary and linguistic training can be used in highly competitive fields of design construction and manufacturing.

The social change that we are expecting from a well-equipped graduate with soft skills would turn into the better professional with more options to move upwards, toward a good career, that would be beneficial for his family and organization both. These skills must be implemented as a compulsory subject in the syllabus of the college education, for the growth of a student’s personality. Finally, in the words of Fyodor Dostoyevsky, “Much unhappiness has come into the world because of bewilderment and things left unsaid” [5].

## References

1. <https://dictionary.cambridge.org/dictionary/eng>
2. [https://www.gradesaver.com/author/robert\\_frost](https://www.gradesaver.com/author/robert_frost)
3. <https://www.thefamouspeople.com/profiles>
4. <https://www.lexico.com/en/definition/softskills>
5. Soft skills in “A dictionary of human resources management”
6. Developing soft skills and personality. Professor Ravichandran, Department of humanities and social sciences, IIT, Kanpur
7. Danah Zohar, SQ (2000) Connecting with our spiritual intelligence, London: Bloomsbury, 2000  
Defined 12 principles underlying spiritual intelligence

# Virtual Machine Migration Approach in Cloud Computing Using Genetic Algorithm



Gursharanjit Kaur and Rajan Sachdeva

**Abstract** The technology of cloud computing is decentralized in nature due to which various issues in the network get raised which reduce its efficiency. The cloud computing technology is applied to fulfill the demands of hosts over the internet. For the purpose of using or sharing, the resource cloud computing can be used. The virtual machine migration is the major issue of cloud computing, and it gets raised when uncertainty get happened in the network. Due to extensive use of the virtual machine resources, machine gets overloaded which increase delay for the cloudlet execution. In the base paper, the threshold algorithm has been proposed which assign task to most capable machine and hosts maintain checkpoints on the virtual machines. When the virtual machine get overloaded, the task needs to migrate to another VM (Virtual Machine). In this study, weight-based technique will be proposed which migrate cloudlet from one virtual machine to another.

**Keywords** Genetic algorithm · Cloud computing · Load balancing

## 1 Introduction

The technology of cloud computing provides the access of the network to various subscribers as per their demand. This technology offers various computing assets, e.g., applications, storage space, etc. This technology can be used for maximizing the efficiency of the cloud [1]. Cloud technology enables the reusability of IT resources for storing large databases, developing and hosting complex applications, and expanding computational power and other services on demand. Eliminating or reducing investments on large-scale infrastructure and software, coupled with the pay-per-use model, significantly reduces IT costs. The client or an institution stores

---

G. Kaur · R. Sachdeva (✉)

Guru Gobind Singh College of Modern Technology, Kharar, Punjab, India

e-mail: [hodcse.ggcmt@gmail.com](mailto:hodcse.ggcmt@gmail.com)

G. Kaur

e-mail: [ergursharan18@gmail.com](mailto:ergursharan18@gmail.com)

extracted and improved data in centralized data. It is known as cloud. In cloud, the service providers offer cloud services to the clients as per the demand. This fundamental trait is called CSP. CSP represents “Cloud Service Provider”. It implies that the customer utilizing the service pays a certain amount of money in return. Cloud computing is a technology that provides a complex number of applications in various scenarios. Every scenario provides some expert particular service. There are basically three types of clouds available. The first one is called Public cloud. This cloud generally is SaaS services. These services are provided to the clients using Internet. It is the most cost-effective alternative for clients. In public cloud, the provider of service tolerates the cost of both bandwidth cost as well as configuration. This cloud has restricted shapes [2]. The potential of clients of using services determines the cost here. However, this approach has several constraints such as deficiency of SLA parameters. This approach offers high consistency, lower price, null preservation, and scalability as per demand. But despite these advantages, this approach cannot be used for those associations that work with responsive information. The designing and development of private cloud are done as per the requirements of an individual company. The service providers of this type of cloud provide services with better security. In this way, they ensure that the outsiders could not get the access of the service [3]. Therefore, in contrast to public cloud, this type of cloud is more safe but less adaptive, e.g., Amazon Virtual Private Cloud. The third type of cloud is known as hybrid cloud. This type of cloud combines private and public cloud. This type of cloud offers more adaptability to an organization. This cloud also controls crucial tasks and resources. These are joined together for improving adaptability and reducing overhead. The methodology of Load balancing can be used to share the network’s load. This approach uses different methods for this purpose. This approach enables effective resource usage and improves the response time of the task [4]. Meanwhile, this approach eliminates a condition having some centers under loaded. At the same time, few hubs are overloaded. This approach makes use of various tools for load balancing rather than a single module. In this way, this approach increases the consistency and accessibility of the data by eliminating redundancy. The computation of load is carried out in terms of figures in the region of processor load, memory usage, network load, etc. A program on a virtual space within a host acts as a single object. This object caters the request and offers services similar to a physical structure. It is named as virtual machine. It behaves as a whole system in spite of being virtual within a host. In general, a VM is generated within a larger atmosphere called as host [5]. A host may contain numerous virtual machines. These machines act as an autonomous object. Genetic algorithm is derived from the research works carried out on cell automata. John Holland and his colleagues developed this algorithm. Originally, GA (genetic algorithm) is one of the popular techniques. This approach is quite useful in the discipline of CSE (computer science). This approach can resolve issues related to optimization. These algorithms are termed as the evolutionary algorithms [6]. In this, numerous methods are included by developmental science, for example, legacy, change, characteristic determination, and recombination. In the representation of the hereditary calculations, the wellness capacity is characterized. The hereditary computation continues to instate the arrangements arbitrarily. It used to enhance it

through monotonous application. In this case, it involves many applications such as selection, mutation, and crossover operators. Numerous researchers have embraced hereditary calculations as an answer for streamlining in different fields. The hereditary calculations goes about as an answer for improvement issue began picking up fame towards the end of the most recent century as used to tackle enhancement issues in development [7]. Its natural parallelism encourages the employments of circulated preparing machines, similar to Distribution Network Planning. Issues which have all the earmarks of being especially fitting for arrangement by GA incorporate Scheduling and State Assignment Problem. To solve color problem, GA has been used many a times which shows efficiency of GA in this matter. Specialists have indicated enthusiasm for GA way to deal with take care of booking sorts of issues, similar to employment shop planning issue. It can be very viable to consolidate GA with other enhancement systems.

## 2 Literature Review

Karki et al. in 2018 [8] stated that the storing of data had been carried out in a centralized VM (virtual machine). It was called cloud. The cloud providers were accountable for the distribution of cloud services to the ultimate customer. The ultimate customers got the access of these services as per their demand. Also, these customers had to pay a certain amount for using these services. It was imperative to balance the load with the increase in the demand. Load balancing could minimize the usage of valuable assets and the expenditure of power. The overloading of VMs (virtual machine) during the implementation of cloud could be reduced by mitigating task. Task migration could be carried out using threshold and check pointing approach. This approach could mitigate tasks from one to another VM or arrange them in queue. This resulted in minimum processing time, power as well as resource usage.

Kaur et al. in 2017 [9] recommended IGA (improved genetic algorithm) approach to balance the load. This approach assigned the chores of clients to the VM (virtual machine). Maximizing the usage of resources and minimizing the energy expenditure and the implementation overhead of task were the key objectives of this approach. IGA gave higher output in terms of energy efficiency; cost and all the VMs distributed the tasks in such a way that the load is properly balanced. The achieved simulation outcomes were represented in the form of graph. These outcomes proved the efficiency of recommended approach over other existing load balancing approaches.

Bei et al. in 2016 [10] recommended a novel load balancing task scheduling approach. The recommended approach was called MPGA (multi-population genetic algorithm). A lot of simulation tests were carried out to test the working of recommended approach in terms of scheduling. This approach made use of min-min and max-min algorithms for initializing the populace. Afterward, this approach employed metropolis measure for the avoidance of local optima. The tested outcomes revealed that the recommended approach outperformed the other existing approaches in terms

of different performance metrics. This approach minimized time consumption as well as processing overhead. This approach also efficiently carried out the load balancing of inner devices.

Pilavare et al. in 2015 [11] studied that the area of cloud computing made use of several approaches for making improvements in the balancing of load. It was analyzed that GA (genetic algorithm) performed most efficiently among all employed approaches. This approach selected virtual machines in random manner. It was used input and after that the processing had been carried out. The earlier approaches assigned similar priority to the tasks and the virtual machines. However, it was not practical. Therefore, at first, the input processors were inserted to the priority algorithm. This algorithm was called Logarithmic Least Square Matrix. This algorithm resolved some issues related to idleness and starvation.

Gai et al. in 2015 [12] recommended a new scheme called CAHCM (Cost-Aware Heterogeneous Cloud Memory Model). The main objective of recommended scheme was to provide highly efficient heterogeneous memory service based on cloud. An approach named 2DA (Dynamic Data Allocation Advance) provided support to the recommended approach. This approach used genetic programming for determining the distribution of data on the memories based on cloud. The recommended algorithm considered a variety of important aspects. The aspects influenced the functioning of different metrics. At last, the testing of the recommended approach had been carried out using for its assessment. The achieved outcomes confirmed the feasibility of this approach as a less expensive approach based on cloud.

Mahalingam et al. in 2015 [13] analyzed that it was very important to balance the load for the good performance of cloud computing. An optimized load balancing approach had been recommended. The main aim of this approach was to allocate the approaching tasks over several VMs (virtual machines). A simulation tool called cloud simulator had been used in this work for performance analysis. The outcomes of recommended approach were compared with the outcomes of other available approaches. Cloud simulation was used as a framework which enables modeling, simulation and applicable on making cloud computing infrastructure self-implied platform which has been used to model data centers, hosts, service brokers, scheduling, and allocation policies. In the future work, the problem may be overcome like deadlock and server overflow.

### 3 Research Methodology

The proposed algorithm is the enhancement in the improved genetic algorithm to reduce execution time for the task migration in cloud computing. The number of migration is reduced by changing the mutation calculation points by which the execution is made faster and more reliable than the existing approach. The genetic algorithm works in three phases, the first phase is the initial population in which execution and failure rate of each virtual machine is taken as input. In the second phase, the cross over values is calculated, and in the last step, the best value is selected from the

multiple values which have least chances of failure. In this work, the enhancement in the improved genetic algorithm is proposed to reduce execution time. In the enhanced improved genetic algorithm following steps are there:-

1. Initial Population: The initial population is the execution time and failure rate of each virtual machine which is used for the task execution. The initial population is the virtual machine resource which is used for the task execution.
2. Cross over value calculation: Under the populace, every chromosome  $\times$  fitness value gets explored. According to their fitness value, we pick two parent chromosomes from a populace, and in general, bigger the populace bigger is the fitness value with a crossover possibility exceeds the parents to form a new offspring. If none of the crossover was done, offspring becomes exact replica of parents. With an offspring, possibility recreates new offspring at every locus.
3. Best value Calculation: The best value is calculated from the crossover value calculated. Use the new generated populace for a farther run of the algorithm (Fig. 1).

## 4 Experimental Results

The implementation of recommended research is carried out in MATLAB tool. The outcomes of recommended and earlier approaches are compared for evaluation purpose with respect to certain performance parameters.

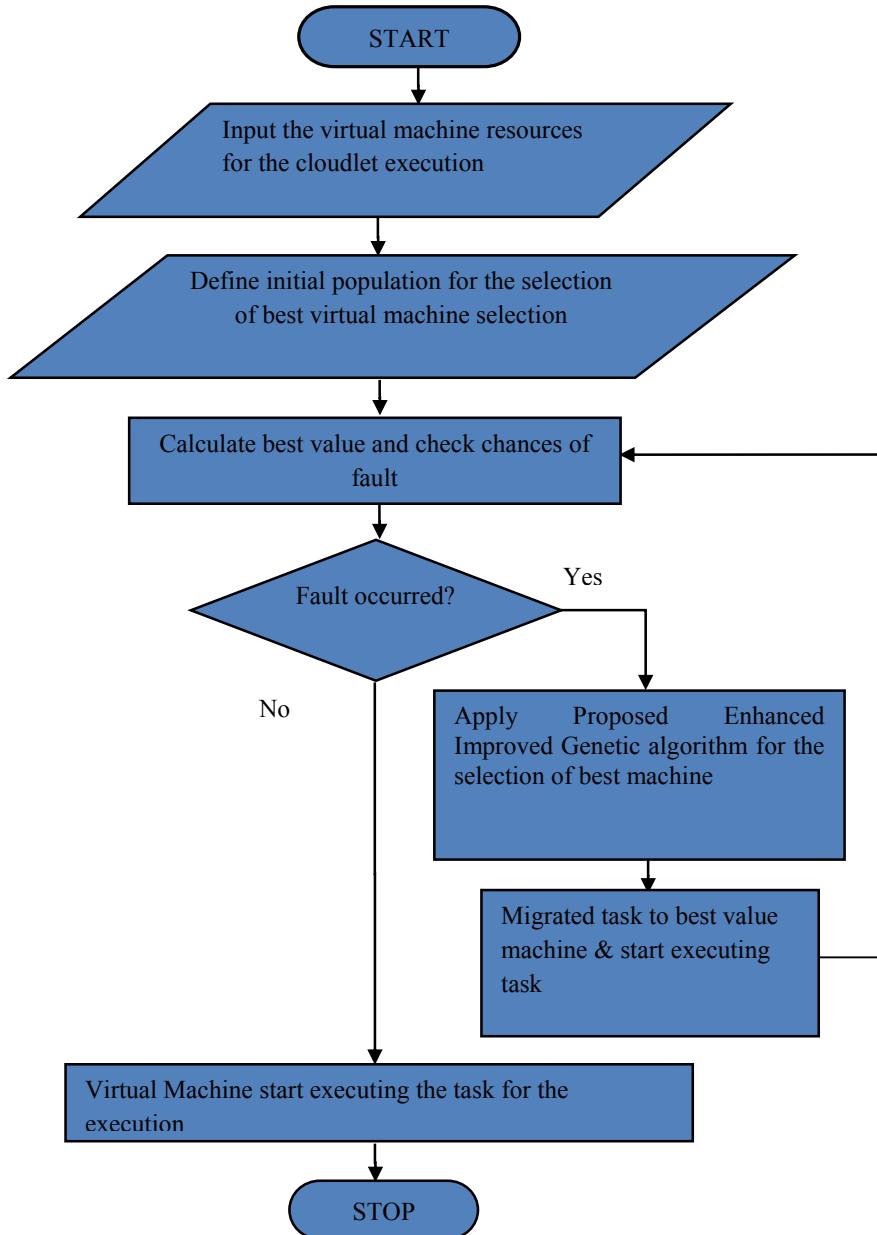
The performance of recommended approach is analyzed by comparing this approach with the improved GA (genetic algorithm) in terms of response time as shown in Fig. 2. The response time of recommended approach is lower than the improved GA (genetic algorithm).

The performance of recommended approach is analyzed by comparing this approach with the improved GA (genetic algorithm) in terms of finish time as shown in Fig. 3. The finish time of recommended approach is lower than the improved GA (genetic algorithm).

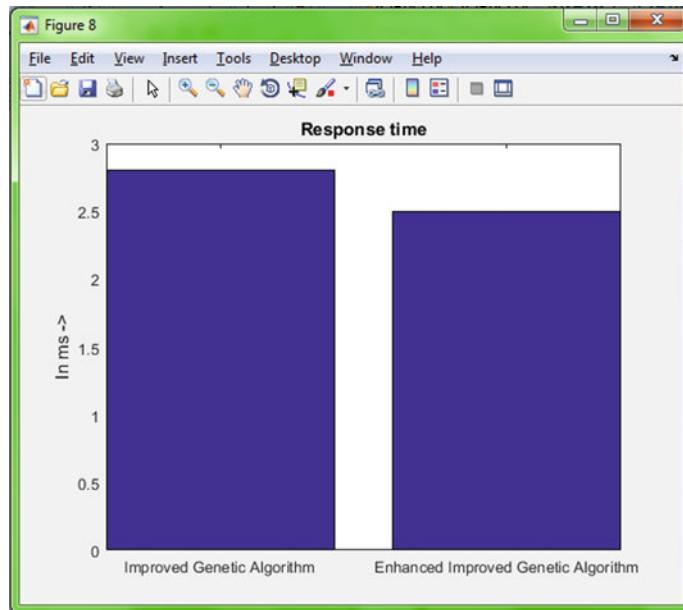
The performance of recommended approach is analyzed by comparing this approach with the improved GA (genetic algorithm) in terms of energy usage (consumption) as shown in Fig. 4. The energy usage of recommended approach is lower than the improved GA (genetic algorithm).

The performance of recommended approach is analyzed by comparing this approach with the improved GA (genetic algorithm) in terms of cost as shown in Fig. 5. The cost of recommended approach is lower than the improved GA (genetic algorithm).

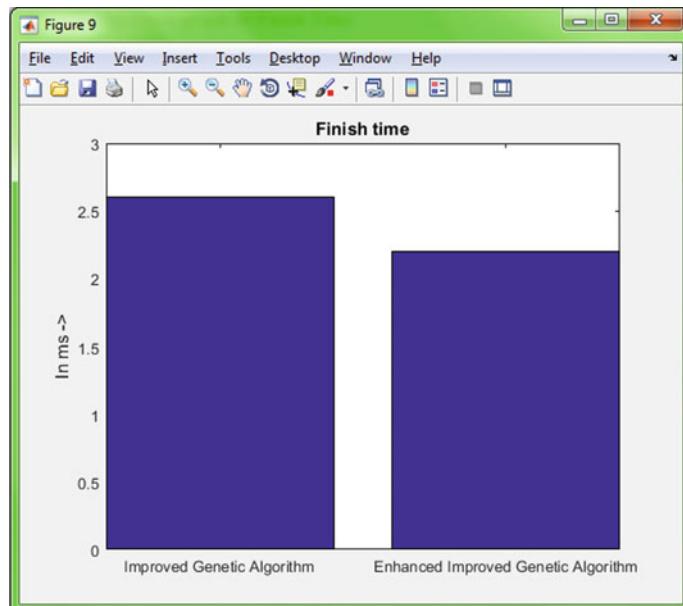
The performance of recommended approach is analyzed by comparing this approach with the improved GA (genetic algorithm) in terms of number of mitigations as shown in Fig. 6. The number of mitigations of recommended approach is lower than the improved GA (genetic algorithm).



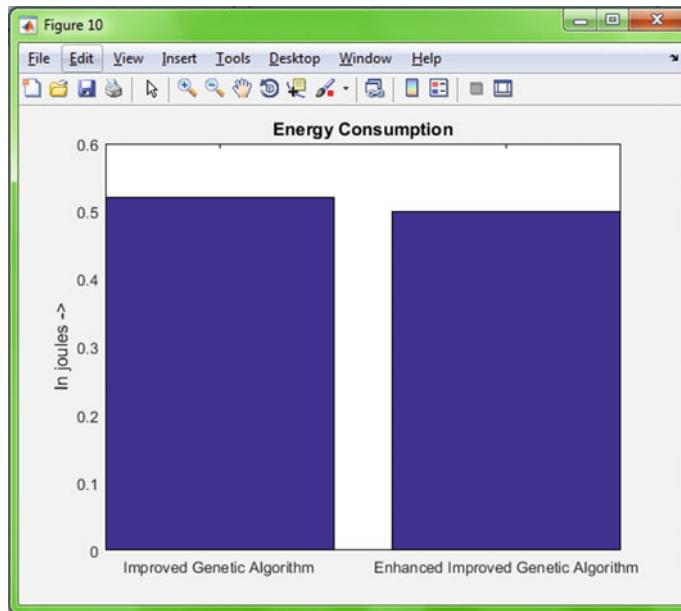
**Fig. 1** Proposed flowchart



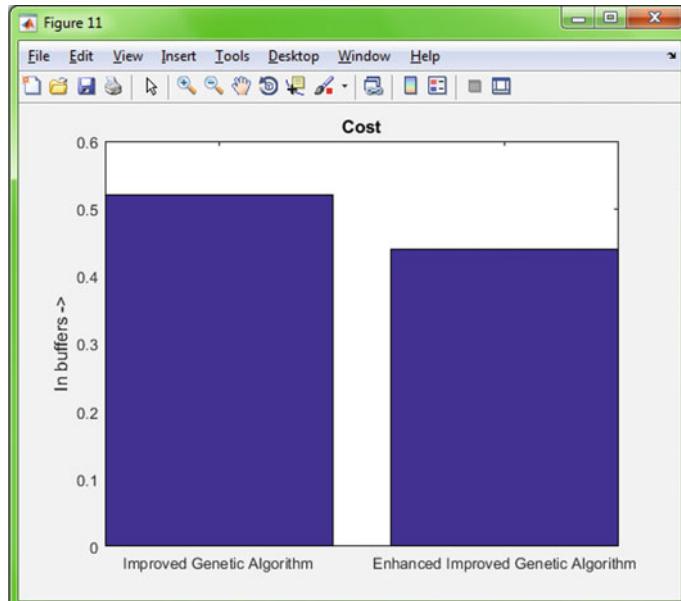
**Fig. 2** Comparison graph of response time



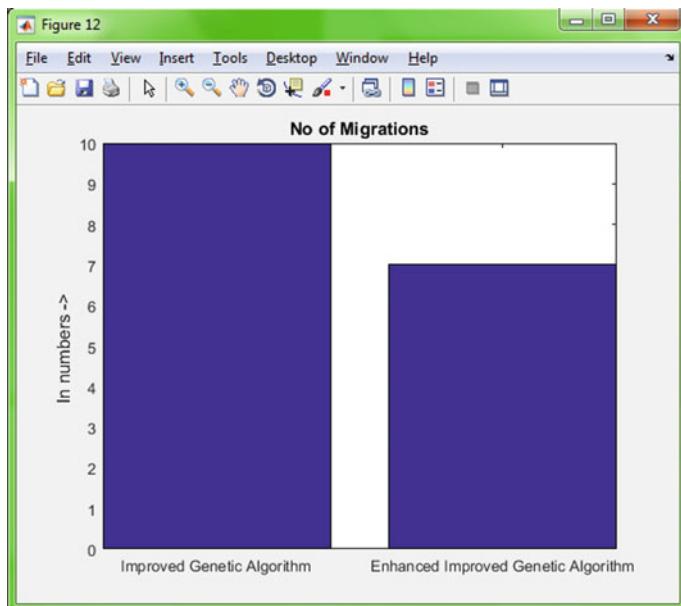
**Fig. 3** Comparison graph of finish time



**Fig. 4** Comparison graph of energy consumption



**Fig. 5** Comparison graph of cost



**Fig. 6** Comparison graph of no of migrations

## 5 Conclusion

The technology of cloud computing faces several problems due to its dynamicity. These concerns include safety, quality of service and fault incidence and so on. The load balancing is one of the key concerns of cloud network. This issue decreases the competence of cloud. In this work, an enhancement in the improved genetic algorithm is recommended for minimizing the execution time. This recommended approach has high consistency and pace. Therefore, this approach minimizes the probabilities of fault incidence. The implementation of recommended and the other available approaches is carried out using MATLAB tool. These approaches were compared for performance evaluation. For this purpose, different matrices have been considered. On the basis of the graphs, it is analyzed that the recommended GA (genetic algorithm) outperforms the other existing improved GAs (genetic algorithm) in terms of the migration of VM (virtual machine).

## References

1. Barron C, Yu H, Zhan J (2013) Cloud computing security case studies and research. In Proceedings of the world congress on engineering 2
2. Guo W, Wang X (2013) A data placement strategy based on genetic algorithm in cloud computing platform. In 10th web information system and application conference

3. Gowri G, Amutha M (2014) Cloud computing applications and their testing methodology. *Int J Innov Res Comput Commun Eng* 2906–2914
4. Gupta R (2014) Review on existing load balancing techniques of cloud computing. *Int J Adv Res Comput Sci Softw Eng* 4(2)
5. Wang T, Liu Z, Chen Y, Xu Y (2014) Load balancing task scheduling based on genetic algorithm in cloud computing. In: IEEE 12th international conference on dependable, autonomic and secure computing
6. Singh A, Juneja D, Malhotra M (2015) Autonomous agent balancing algorithm in cloud computing. *Procedia Computer Science International Conference on Advanced Computing Technologies and Applications (ICACTA)* 45, 832–841 (2015)
7. Wang J, Zhao Z, Xu Z, Hu Z, Li L, Guo Y (2015) I-sieve: An inline high performance deduplication system used in cloud storage. *Tsinghua Sci Technol* 17–27
8. Karki S, Goyal A (2018) Performance evaluation of check pointing and threshold algorithm for load balancing in cloud computing. *Int J Comput Sci Eng* 6(5):2347–2693
9. Kaur S, Sengupta J (2017) Load balancing using improved genetic algorithm(iga) in cloud computing. *Int J Adv Res Comput Eng Technol (IJARCET)* 6(8), 2278–1323
10. Bei W, Li J (2016) Load balancing task scheduling based on multi-population genetic algorithm in cloud computing. In: Proceedings of the 35th chinese control conference 27–29
11. Pilavare MS, Desai A (2015) A novel approach towards improving performance of load balancing using genetic algorithm in cloud computing. In: IEEE sponsored 2nd international conference on innovations in information embedded and communication systems, ICIECS'15
12. Gai K, Qiu M, Zhao H (2015) Cost-aware multimedia data allocation for heterogeneous memory using genetic algorithm in cloud computing. *IEEE Trans Cloud Comput*
13. Mahalingam NN (2015) Efficient load balancing in cloud computing using weighted throttled algorithm. *Int J Innov Res Comput Commun Eng* 3:5409–5415

# A Survey on Electronic Health Records Using Cloud Computing Environment



Vivek Gehlot, S. P. Singh, and Akash Saxena

**Abstract** One of the serious problems existing in the cloud is to oversee or to verify the information's as of unapproved people. Here in the restorative field, the patient-driven model depicts the patient's personal health record, wherever health data is to be verified from the outside servers. Electronic health record difficulties incorporate expensive programming bundles, framework security, tolerant secrecy, and obscure prospect government guidelines. Impending advancements via electronic health records incorporate bar coding, radio-recurrence ID, also discourse acknowledgment. We have provided a study on electronic health records in cloud computing with various technical aspects. Also, there is a detailed description of their applications. Load balancing is also the most significant aspect of cloud computing.

**Keywords** Cloud computing · Electronic health records · Load balancing · Technical aspects

## 1 Introduction

Cloud computing (CC) or the eventual fate of cutting edge registering furnishes its customers with a virtualized system access to applications as well as administrations. Regardless of from any place the customer is getting to the administration, he is consequently coordinated to accessible assets. In some cases, our framework becomes draped up or it appears to yield a couple of times for pages to leave the

---

V. Gehlot (✉)

Research Scholar, Department of Computer Science and Engineering, Nims Institute of Engineering and Technology, Nims University Rajasthan, Jaipur, India  
e-mail: [vivekgehlot369@yahoo.com](mailto:vivekgehlot369@yahoo.com)

S. P. Singh

Professor and Head, Department of Computer Science and Engineering, Nims Institute of Engineering and Technology, Nims University Rajasthan, Jaipur, India

A. Saxena

Professor, Department of Computer Science, Compucom Institute of Information Technology and Management, Jaipur, India

printer. This occurs in light of fact that there is a line of solicitations trusting that their turn will get to assets which are shared among them. Be that as it may, these solicitations can't be overhauled as assets obligatory by every one of these solicitations are detained by a new procedure or solicitations by virtual machines. One reason for every one of these issues is named halt. LB is another methodology that helps systems and assets by giving the great amount as well as least reaction time [1]. In cloud stages, asset designation (or burden balancing) happens dominant part at two stages.

1. At first level: The LB allots mentioned occurrences to physical PCs at the season of transferring a use endeavoring to adjust the computational heap of many applications transverse over physical PCs.
2. At second level: Once an application becomes various approaching solicitations, every one of these solicitations must be doled out to a particular application occasion to regulate computational burden concluded a lot of samples of a similar application [2].

Today, the progression of innovation uniquely is therapeutic sciences has transformed human services associations into client situated conditions [3, 4]. These associations are in mission for superiority development. This won't be accomplished with no time admittance to top-notch data [5]. As per the meanings of International Organization for Standardization (ISO) and Electronic Health Record (EHR) are capacity, safe trade as well as access to persistent data in computerized design by few approved clients. This data incorporates patients past, present as well as future data. The principle target of EHR is to help support of coordinated, effective, and quality health [6]. In other definition, EHR incorporates all data identified with the soundness of natives from before birth (data about prenatal and postnatal hatchlings, for example, in vitro treatment data) to after death (data got from an examination, landfill and so on). This data put away consistently and electronically after some time. On the off chance that vital, without contact with the area or time, all or some portion of this data will be accessible to approved people [7]. For the most part, coordinated EHR partners are individuals from the general public. All social insurance suppliers are partners and clients. EHR has a critical change in giving medicinal services, decreasing mistakes, and expanding the effect of health care [8]. Straightforward entry to all data on patient history improves care, centers around data and along these lines decreases therapeutic symptomatic mistakes. The synchronous accessibility of particular restorative focuses on the EHR is a significant advantage. Additionally, it is imperative to keeping up and insurance the data from deterioration, breakdown, and obliteration under any circumstances is another.

## 2 Load Balancing

Load balancing (LB) is characterized as the dispersion of assets, concurrent working of schedulers, proficiency upgrade, and minimization of reaction time by means of a

reasonable coordinating of employment to the accessible asset. Concurrent working of the schedulers includes circulation of burden in an equivalent way among processors. To reestablish parity dynamic burden balancing otherwise called burden-sharing or burden relocation is utilized [9]. It is finished by appropriating the whole burden to separate mainframes of total construction aimed at acquiring productive asset mapping also simultaneously evacuating the likelihood of over-burdening or under stacking of hubs in system. It is an improved proportion of client acknowledgment and asset use, along these lines upgrading the throughput of the total framework. Whenever done in a legitimate way heap board can constrain the utilization of accessible. It additionally helps in executing disappointments, making the framework adaptable and over-loading, limiting reaction time and so on. The primary objectives of LB calculations are recorded underneath [10]:

1. Rate efficiency: Load offsetting backing ought to perform by a lesser rate in the assumed system. Toward create positive unnecessary proficiency, diminished reaction time, and keep the over-burden.
2. Scalability: Configuration via which LB strategies are performed ought to be equipped for being modified in components in a future period.
3. Elasticity: Down to earth prerequisites ought to have the option to make do with conditions; they ought to be flexible and versatile. To force prospect varieties in the framework.
4. Priority: The assets and errands ought to be orchestrated according to need. Thus, advanced working gives indications of redesigning as well as implementation.
5. Source usage: To ensure that accessible assets are being used in an ideal way.
6. Backup: If there should arise an occurrence of the disappointment of the structure, LB calculations need to reinforcement plan.
7. Homogeneous nature: To extravagance all undertakings in framework homogeneously independent of their source [11, 12].

### 3 Bases of Cloud Computing: Technical Aspects or Challenges

There are numerous specialized problems in CC that should be moved before these benefits may be entirely recognized, which include foundation, load balancing, security also protection in CC as well as so on. Among them, LB is an essential scheme to magnify service level agreement (SLA) as well as better employment of assets.

- A. **Infrastructure.** Cloud supplier wants to deal through all apparatus as well as scheme to provide better management to termination client. On-off accidental that problem in a foundation which advance issues in giving management similar SaaS, then a bunch in the cloud may get lopsided since of poor framework. This prompts poor QoS [13].

- B. **Load Balancing in Cloud Computing.** A central point must be handled in CC is LB, numerous variables similar deprived foundation, awful traffic board, establishes resolute class reminders, lopsided group. In little schemes, it tends to be unrelated however in a complex scheme, to provide better directions all these are main consideration to fare the well whereas structuring algorithm aimed at multifaceted scheme [13].
- C. **Security and Privacy in Cloud Computing.** Termination clients supply their information on bases of safety as well as protection in the cloud yet because of numerous reasons like the development of information also an application on the system, loss of regulator proceeding information, assaults on information and so on. The security may get affected. To perceive this issue is a real test in CC [14].
- D. **Trust in Cloud Computing.** At the point after customer or termination client demand an administration as of cloud, there is administration level understanding wants to sign or requirements to concede to terms also a state of cloud supplier. This has absolutely relied upon trust of customer proceeding cloud supplier. Belief is the all-encompassing type of safety as well as protection. Two kinds of belief are characterized in [14].
  - 1. Hard trust (security-arranged) in view of legitimacy, programming as well as safety in.
  - 2. Soft trust (non-security arranged) in view of human brain science, steadfastness to exchange mark (brand reliability), and ease of use.
- E. **Ensuring Data Portability and Interoperability.** There need to information versatility in CC, similar to the capacity to alter sellers, later on, organizations can attempt to avoid phases or innovations that “lock” clients in a precise item 7.

## 4 Electronic Health Record

Electronic health records (EHRs) are progressively being sent advanced inpatient data frameworks of clinical and authoritative information. Transaction of social and specialized variables is significant after considering successful usage and appropriation procedures in occupied complex medical clinic conditions. These frameworks offer significant potential to improve health, quality, and productivity of clinic medicinal services arrangement; however, understanding these advantages is vigorously reliant on framework streamlining. The improvement of inpatient data frameworks is best conceptualized as a continuous voyage. Advancements, for example, transition to cloud-based EHRs, opening up of utilization program interfaces, chance to associate with other digitized clinic framework, for example, savvy mixture siphons and beds, understanding access to EHRs and improvements in ways to deal with and limit with regards to cross-examining independent and connected EHR-based

datasets progressively present major new chances to improve results. These improvements will anyway likewise bring significant new moral, authoritative and protection related difficulties that society should address [15].

## 5 Electronic Health Record Applications

- A. **Administrative Applications.** EHR must have a specific degree of authoritative applications. Regulatory use is a piece of EHR that incorporates persistent enrollment. Enduring enlistment is placing persistent socioeconomics are verified on health record (HR); also, this incorporates the name, age, sex, address, contact data, protection data, business, and patient's central objection [16]. The enrollment framework doles out the patient a one of a kind enduring ID no. that is just utilized through specific medicinal services supplier.
- B. **Computerized Physician Order Entry.** Prerequisite for entirely EHR's is an application named electronic doctor request passage or CPOE. CPOE is an application utilized by doctors toward arranges research facility, drug store, radiology administrations, and other doctor orders. CPOE holds incredible points of interest to social insurance suppliers by enabling doctors to electronically request tests without composing these requests on paper frames. This guarantees the precision of the requests and informs the suitable region that the patient will arrive. It additionally tells medicinal services experts what tests should be performed. CPOE capacities are additionally an arrangement of the administration's important use necessities [17].
- C. **Laboratory Systems.** Most research services in medicinal services settings as of now usage lab information system (LIS) which is normally interfaced in EHR aimed at patient information as well as challenging outcomes trade. Practically entirely lab analyzers as well as lab testing gear interface in LIS. LIS additionally comprises lab orders, lab outcomes, plans as well as new regulatory capacities [16].
- D. **Radiology Systems.** Radiology information system (RIS) is alternative office by data frameworks that edge by EHR. RIS, similar to the lab system, contains tolerant data, radiology orders, test outcomes, calendars also picture following. RIS additionally utilizes with picture achieving communication system (PACS). This is a framework that oversees and then supplies computerized radiography picture. Computerized radiology pictures can be shared to see inside EHR application [16].
- E. **Clinical Documentation.** Clinical documentation is a massive part of an EHR, doctors, attendants as well as new medicinal services authorities' noise massive measure of information on persevering. These data varieties as of clinical note's clinical reports, appraisals as well as medication administration records (MAR). Dissimilar shares of clinical certification include imperative symbols, issue outlines, interpretation reports as well as usage executives [18].

- F. **Pharmacy Systems.** Pharmacies in enormous clinics are much computerized, applying automata to fill solutions as well as utilize electronically incorporated prescription automobiles. These independent drug store frameworks are an alternative framework that is interfaced through an EHR. Emergency clinic drug stores additionally use bar coding on meds as well as patients to assurance right helping, right patient, perfect period, medicate organization. It is crucially significant that drug store frameworks interface in EHR as this is place sedate associations too medication sensitivities are shadowed privileged an EHR. Medication errors in human services are a key source of restorative blunders that goal persistent damage. An important segment of EHR's drug store applications is e-recommending [19]. Unique wellspring of medicine blunders is a specialist's messy penmanship proceeding solutions and medication orders. E-recommending wipes out this issue by sending solution electronically to a retail drug store or emergency clinic's drug store. EHR's are an incredible apparatus to help diminish or dispose of medication mistakes.
- G. **Other Applications.** Numerous EHR's comprise different applications that support kind of progressively whole record. A significant application is scientific choice help [20]. Clinical choice emotionally supportive networks help doctors and medical attendants to pick the right strategy on specific persistent and his/her state. Alternative significant use that is a piece of important usage guidelines is superiority administration frameworks. Quality administration frameworks track tolerant results and give human services supplier's devices to report the information to government elements.

## 6 Literature Review

Sarwar et al. (2019) a cloud-based engineering for the usage of EHR framework for emergency clinics of Pakistan is proposed. Receiving proposed cloud-based design will be useful in improving patient consideration, diagnostics, malady introduction, and then nonstop accessibility of EH data. In addition, EHR framework will help in decreasing the expense of keeping up paper-based records. Advancement of such framework won't just empower specialists and clinics to trade patient's data with one another, however, will likewise build up electronic health information store that therefore can be utilized for differing purposes, for example, prescient diagnostics and customized drug [21].

Joshi et al. (2018) built up a novel, incorporated, possession created support scheme that utilizes attribute-based encryption (ABE), then proceeds in consideration designated safe admittance of enduring archives. This component moves administration board overhead from patient to medicinal connotation and then certifies simple designation of cloud-based EHR's appearance specialist to restorative suppliers. In this paper, we depict this novel ABE method just as a model framework that we consume makes to delineate it [22].

Manoj et al. (2017) another safe mixture EHR is suggested. In this structure, two proficient encryption strategies are joined for acceptable grained admittance control as well as insurance of data protection. Multi-specialist also key-based encryption plans are utilized for encryption of every piece of HR subsequent to isolating those archives using perpendicular dividing strategy. Multi-expert encryption plans are basically utilized in public domains (PUDs), though key-based encryption plans are common in personal domains (PSDs). Composed, they give safe information admittance also confirmation of clients. Execution is encouraged utilizing Windows Azure CC stage [23].

Bansal et al. (2016) working over upgrading the safety of PHR through usage of DWT that is created ended steganography. Steganography is that procedure where data can be hidden in an image. Steganography has a capacity of embedding new data in one picture. PHR information will be hidden in a unique picture through steganography. In steganography, encryption and decoding procedure will be performed via DWT. In this paper, we are working for concealing PHR information in the solitary picture for security reasons. For concealing PHR information in the picture, wavelet created steganography is suggested in this paper. In outcomes, elapsed period then error rate get lessens up to 80–90% [24].

Hameed et al. (2015) suggested a model of preparation flexible e-human facilities executive's structure dependent on CC, then service-oriented architecture (SOA). Cloud and SOA are getting to be omnipresent these days. Notwithstanding, their applications are sent around there and wanted to seem different period stringent additionally useful conditions. An optional outline has been better as well as includes dissimilar portions to generate a medicinal services structure. Rich internet application (RIA) in light of customer side, basic folder cloud server as well as application side prompts accomplishing a plan that makes promptly achievable system. At long last, the structure suggested improves price executives, period, knocking away patient's summary, and then taking correct specialist choice [25].

Bahrami and Singhal (2015) present a novel CC stage dependent on service-oriented cloud design. A proposed stage can be kept running on the highest point of heterogeneous cloud computing frameworks that gives standard, dynamic and adaptable administrations for eHealth frameworks. The proposed stage permits heterogeneous mists to give uniform administration interface to eHealth frameworks that empower clients to unreservedly move their data and application starting with one merchant then onto next with insignificant alterations. We execute a proposed stage for an eHealth framework that keeps up patient's information protection in the cloud. We consider an information openness situation with executing two techniques, AES then light-weight information security strategy to ensure patient's information protection on the proposed stage. We evaluate the presentation and adaptability of the actualized stage for an enormous electronic restorative record. The test results demonstrate that the proposed stage has not present extra overheads when we run information security insurance techniques on the proposed stage [26].

Yan et al. (2014) in request to guarantee the safety of information, anybody may not peruse as well as alter it without a grant. In this paper, we usage ABE plot which is an augmentation of identity-based encryption (IBE) to build framework. In this

framework, we receive key-arrangement characteristic based encryption (KACBE) which may gives fine-grained get to strategy, then better proficiency aimed at client renouncement [27].

Khan and Bai (2013) expect to build up a methodology that empowers CC customers to check health administrative consistence asserted by CC suppliers. In this methodology, customers of CC squared consequently how cloud provider happens administrative consistence, for example, HIPAA enactment for their health records. In spite of the fact that cloud suppliers frequently outfit their administrations with outsider accreditations on gathering administrative compliances, the customer does not have any way to confirm how administrative compliances are really accomplished in a wide assortment of cloud administration situations in connection to their electronic secured health data (e-PHI). Our methodology depends on three procedures: (i) Mechanisms to speak to health guidelines in the machine processable structure; (ii) Collection of administration explicit consistency related to continuous information after cloud servers; and then (iii) Involuntary thinking around compliances amongst machine processable guidelines as well as gathered information as of servers [28].

Bahga and Madisetti (2013) propose an EHR framework—cloud health information system technology architecture (CHISTAR) that accomplishes semantic interoperability using conventional plan philosophy which usages orientation model that characterizes universally useful arrangement of information structures and paradigm model that characterizes the clinical information properties. CHISTAR application segments are planned to utilize cloud segment model methodology that involves approximately coupled parts that impart no concurrently. In this paper, we depict an abnormal state structure of CHISTAR and then methodologies aimed at semantic interoperability, information combination and then safety [29].

## 7 Conclusion

Expanding the scope of new and regularly exceptionally complex advancements is being utilized in inpatient settings and appropriation of EHRs is presently unavoidable. In any case, usage and reception of frameworks are confounded by hierarchical intricacy of emergency clinics and different social outcomes regularly connected with mechanical change activities. Personal health record (PHR) is a basic health-related administration framework to save health information for somebody. These certainties are put away on untrusted servers which make secure information sharing system. The procedure of cloud servers is ordinarily accessed by third user; generally, PHR arrangements are done through on these clouds. Major encryption strategies are utilized can be completely encoded public health records are put away on capacity cloud.

## References

1. Shimonski R (2003) Windows 2000 and windows server 2003, clustering and load balancing. McGraw-Hill Professional Publishing, Emeryville, CA
2. Tai J, Zhang J, Li J, Meleis W, Mi N (2011) A R A: adaptive resource allocation for cloud computing environments under bursty workloads. IEEE
3. Wing P, Langelier M, Contenelli T, Armstrong D (2003) Data for decisions: the him workforce and workplace—2002 member survey. American Health Information Management Association, Chicago
4. Mon DT (2009) American Health Information Management Association (AHIMA) written and oral testimony at the NCVHS privacy, confidentiality and security subcommittee hearing on personal health records. <https://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/090520p06.pdf>
5. Sittig DF, Singh H (2011) Defining health information technology-related errors: new developments since to err is human. Arch Intern Med 171(14):1281–1284
6. Schloeffel P (2002) Electronic health record definition, scope and context
7. Open HER, What is Open HER ? [https://www.openehr.org/what\\_is\\_openehr](https://www.openehr.org/what_is_openehr)
8. Riazi H, Fathirosari B, Bitaraf E (2007) Electronic health record concepts, standards and methodology, ministry of health and medical education center for information and statistics management, Tehran
9. Jain A, Kumar R (2016) A comparative analysis of task scheduling approach for a cloud environment. In: 3rd international conference on computing for sustainable global development, pp 1787–1792
10. Florence AP, Shanthi V (2013) Intelligent dynamic load balancing approach for computational cloud. Int J Comput Appl 15–18
11. Alakeel AM (2010) A guide to dynamic load balancing in cloud computing systems. Int J Comput Sci Inf Sec 10(6):153–160
12. Jain A, Kumar R (2017) Critical analysis of load balancing strategies for the cloud environment. Int J Commun Netw Cloud Syst 213–234
13. Kumar V, Prakash S (2014) A load balancing based cloud computing techniques and challenges. Int J Sci Res Manage 2(5):815–824
14. Hashemi S (2013) Cloud computing technology: security and challenges. Int J Sec Privacy Trust Manage 2(5)
15. Cresswell KM, Sheikh A (2017) Inpatient clinical information systems. Key Adv Clin Inf 13–29
16. Seymour T, Frantsvog D, Graeber T (2012) Electronic health records. Am J Health Sci 3(3):201–210
17. Jarousse L (2010) What you need to know about meaningful use. H&HN: Hospitals and Health Networks 84(10)
18. Electronic health records overview. National Institutes of Health (2006) <https://www.ncrr.nih.gov/publications/informatics/EHR.pdf>
19. Roman L (2009) Combined EMR, EHR and PHR manage data for better health. Drug Store News 31(9):40–78
20. Helton J, Lingabeer J, Fraine JD, Hsu C (2012) Do EHR investments lead to lower staffing levels ? Healthcare Fin Manag Assoc 66(2):54–60
21. Sarwar MA, Bashir T, Shahzad O, Abbas A (2019) Cloud-based architecture to implement electronic health record system in Pakistan. IT Professional 21(3):49–54
22. Joshi M, Joshi K, Finin T (2018) Attribute-based encryption for secure access to cloud-based EHR systems. In: 11th IEEE international conference on cloud computing, pp 932–935, San Francisco
23. Manoj R, Alsadoon A, Prasad PWC, Costadopoulos N, Ali S (2017) Hybrid secure and scalable electronic health record sharing in hybrid cloud. In: 5th IEEE international conference on mobile cloud computing, services and engineering, San Francisco, pp 185–190

24. Bansal P, Sharma B, Saxena M (2016) Low error rate based secure sharing of personal health record in cloud computing using DWT Steganography. In: 8th international conference on computational intelligence and communication networks, Tehri, pp 428–431
25. Hameed RT, Mohamad OA, Hamid OT, Tapus N (2015) Design of e-healthcare management system based on cloud and service-oriented architecture. In: E-health and bioengineering conference, Iasi
26. Bahrami M, Singhal M (2015) A dynamic cloud computing platform for ehealth systems. In: 17th international conference on e-health networking, application & services, Boston, pp 435–438
27. Yan H, Li X, Li J (2014) Secure personal health record system with attribute-based encryption in cloud computing. In: 9th international conference on P2P, parallel, grid, cloud and internet computing, Guangdong, pp 329–332
28. Khan KM, Bai Y (2013) Automatic verification of health regulatory compliance in cloud computing. In: 15th IEEE international conference on e-health networking, applications and services, Lisbon, pp 719–721
29. Bahga A, Madisetti VK (2013) A cloud-based approach for interoperable electronic health records. *IEEE J Biomed Health Inf* 17(5):894–906

# IoT Security Architecture with TEA for DoS Attacks Prevention



Vishal Sharma and Anand Sharma

**Abstract** The Internet of Things (IoT) is important in now a day's development of wireless networks enabling to acquire information from the environment, devices, and machines. A number of applications have been implemented in various kinds of technologies. IoT has high coverage to security attacks and threats. There are a number of requirements in terms of security. Confidentiality is one of the main concerns in the wireless network. Integrity and availability are key matters along with the confidentiality. This research focuses on identifying the DoS attacks that can occur in IoT. This paper uses Tiny Encryption Algorithm (TEA) to address these above mentioned security matters. From the algorithms, our proposed solutions can control DoS attack on IoT and any other networks of small devices.

**Keywords** DoS attack · IoT · Security · TEA

## 1 Introduction

As we know that the Internet of Things (IoT) is a new standard that is hastily growing in the area of wireless communications [1], we can be defined IoT as it make connection between daily life's things with each other automatically to complete a new task [2].

The time-bound data like temperature, pressure, etc. are gathering using the IoT devices. Using these collected data, we are able to observe and control of all types of devices in the network.

In the context of these benefits of IoT, this method is used in different fields of smart building, environmental monitoring, smart city, and healthcare monitoring

---

V. Sharma (✉) · A. Sharma

CSE, SET, Mody University of Science and Technology, Lakshmangarh, India  
e-mail: [er.vishu1983@gmail.com](mailto:er.vishu1983@gmail.com)

A. Sharma  
e-mail: [anand\\_glee@yahoo.co.in](mailto:anand_glee@yahoo.co.in)

system and Smart parking [3–6]. Nowadays, in several countries the IoT devices are emerging [7] and these numbers of IoT devices are increasing very rapidly.

Because of wide range of applications, elements in IoT interact via broadcasting messages, which form the messages' broadcasting efficiently. This type of IoT network prone to attacks, where attacker to interfere the networks. It will be easier for attacker to interrupt, fabricate, or even steal the data in the networks that might be the great-secrecy or private confidential information.

Due to the network attack on IoT, the personal and industries' information will get excessive loss. As stated by [8] that uncovered several network attacks as, selective forward, DoS attack, information of bogus routing-based data and snooping by using new Internet of Things structure. Among the various type of network attack, DOS attack is a mostly used attack [9]. Because of the various hazard of network attack, the study of security problem in IoT organization is very necessary.

Hence, the safety requirements are the main distress regarding these harms. One of the key problems is the leak of private information that shows the harm of confidentiality. The integrity issues come over when there is a theft of data and identity. Attacks on integrity can restrict the sensing and control the information. Availability is another inordinate target for the attacks.

This research focuses on denial-of-service (DoS) attack that is high possibility to occur when it comes to availability perspective. DoS attack arises when the system or service that is required cannot be retrieved. Hence, securing broadcast communication protocol is conducted into research. This research proposes an IoT security architecture using a light weight cryptographic algorithm as Tiny Encryption Algorithm.

Section 2 presents the DoS attack and its types, Sect. 3 presents related works for IoT security, Sect. 4 presents the proposed solution, TEA algorithm and its working.

## 2 DoS Attacks

A DoS attack can be used to affect the network connection by creating it unreachable to its users. It is used for flooding the goal with continuous increase in traffic, or transmitting information to causes the crash [10].

The DoS attack is mostly widespread network attack techniques in safety and security of the sensor network. The recurrent targets of DoS attack are web-based servers of prominent bodies such as commerce, media, and banking firms.

DoS Attack types are following.

## 2.1 *SYN Flood*

In SYN flood an attacker sends a bunch of SYN requests to the system which is to be targeted, trying to use prodigious quantities of resources of server to mold the system impulsive to genuine traffic.

## 2.2 *Internet Control Message Protocol Flooding*

It is used for IP diagnostic, errors, and operations in a connectionless protocol. Transmitting the bunch of all types of packets of ICMP to the target which try to process each incoming ICMP request, creating denial-of-service state.

## 2.3 *Teardrop Attacks*

It includes the hacker which is sending incomplete, disorganized, and broken oversized IP fragments and overlapping the payloads to the target's machine. This will lead to crash the servers because of a bug in communicating channel fragmentation is re-assembled.

## 2.4 *Peer-to-Peer Attacks*

It is a kind of distribute network in which different nodes, called "peers," act as transmitter and receivers.

## 2.5 *Low-Rate Denial-of-Service Attacks*

The LDoS attack is designed to implement the retransmit time-out-based strategy to decrease throughput of TCP by manipulating slow-time-scale of TCP.

# 3 IoT Security

The requirement for the lightweight cryptography has been broadly deliberated [11–13], and the limitations of the Internet of Things in terms of controlled devices will be emphasized.

At present there are few selected trivial algorithms of cryptography that do not always manipulate security and its efficiency interchanges. Among the hash functions, block-based cipher, and stream-based cipher, the block-based ciphers present significantly better performances.

mCrypton, is suggested block cipher in [14]. This cipher runs using the option of 8 bytes, 12 bytes, and 16 bytes key size. This algorithmic procedure design is trailed by Crypton [15] though functionality simplification toward increase its performance for the constrained hardware. In [16] which are proposed as Hummingbird-2(HB-2) it was inheritor of Hummingbird-1 [17], in which 8 bytes of initialization vector and 12 bytes (equivalent bits) of key is tested to remain unaffected by all of the former identified attacks. Even the main emphasizes of the weaknesses of the algorithm by the cryptanalysis of HB-2 [18] and according to this the initial key can be upgraded. In [19], we have studied energy consumption of various encryption algorithms including RC4, RC5, and IDEA. The cost of the RC5, RC4 [20], and IDEA [21] ciphers on divergent platforms is deliberated. Though, present algorithms were neglect throughout the analysis.

## 4 IoT Security Architecture of DoS

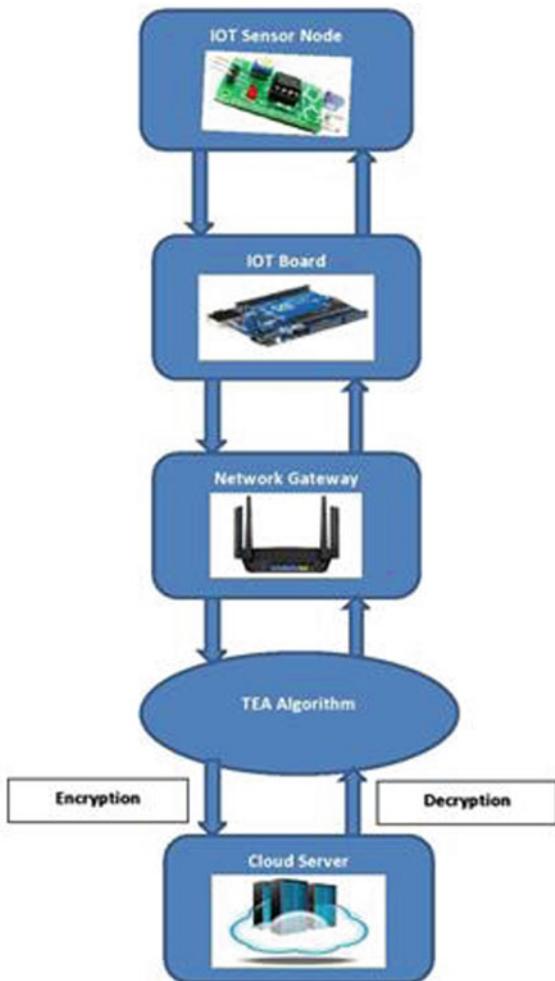
This section is describing the proposed model for IoT security with TEA.

### 4.1 IoT Sensor Node

It is a node in the network of sensor that is capable of acting out some processing. The function of this node is to collecting the sensor's information, and it communicates with various sensor nodes in the sensor network. Sensors of the IoT sensor nodes try to capture data by the surroundings. Sensor nodes are hardware devices that yield a change in a physical condition like temperature, pressure, etc. by quantifiable response. The sensor node got the specific characteristics such as accuracy, sensitivity, etc. by measuring and monitoring the physical data (Fig. 1).

### 4.2 IoT Board

It is a board which has low-power processors which support various programming techniques. The sensor's data is transferred on the cloud server which is collected by various technologies like Wi-Fi, Ethernet, etc. There are different IoT powerful hardware prototyping solutions like Raspberry Pi, Beagle Board, Arduino Uno, Photon, etc. We are using Raspberry Pi 3 for this (Table 1).



**Fig. 1** Implementation of TEA on IoT

#### 4.3 Network Gateway

The network gateway makes the connection between the cloud server and IoT devices. It is used as a physical device or as a software program. All data moving to the cloud and coming from the cloud goes through the gateway. The data which is generated by the IoT devices is preprocessed by the gateway locally at the network devices before sending the data on to the cloud server. This device manages the information moving in either directions, and it can secure the data leaking problem when data is sending to the cloud and IoT devices from being cooperated by malicious attacks by alter detection, encryption.

**Table 1** Technical specification of Raspberry Pi 3

Components	Specification
Central processing Unit	Quad-core 8 byte(64-bit) ARM Cortex A53 clocked at speed 1.2 GHz
Graphical Process Unit	Multimedia 400 Mega Hertz Video Core IV
Memory	1GigaByte LPDDR2-900 SD RAM
Used USB ports	Four
Video Display	High Definition MI, Composite video (NTSC and PAL) Through 3.5 mm jack
Network Supported	10/100 Megabps Ethernet and Wireless LAN(802.11n)
Other Peripherals	HAT ID bus and 17 General purpose IO with specific functions,
Bluetooth support	4.1
Power	5 V using Micro USB/GPIO header
Size	85.60 mm × 56.50 mm
Weight	1.6 ozss

#### 4.4 *Cloud Server*

It is a typical type virtual server that is extracted by the users through a cloud computing platform over the Internet. It grasp and represent the related abilities and functionality like a physical server and from a cloud service provider, it is retrieved remotely.

### 5 TEA Algorithm

This section consisting of two parts, first is describing the introduction of TEA and second will describe the working of TEA.

#### 5.1 *Introduction*

The Tiny Encryption Algorithm (TEA) is simply a block-based cipher given by David Wheeler and Roger Needham. The TEA is a Feistel structure which uses 64 cycles. TEA functions based on two 32 bit unsigned integer and a 128 bit key. The dual shift method used to combine every bits of the data and the key repetitively. In the process of schedule algorithm of the key, the 128 bits of key “e” is divided into the 32-bit blocks as  $E[j]$  equal to ( $j = 0, 1, 2, 3$ ).

```

void ENcode(long* d, long* E) {
    unsigned long s = d[0], t = d1], summ = 0,
    /* setup */
    delta= 0x9e3779b9 /* Delta which is called as a key schedule constant */
    n1= 32;
    while (n1-->0) { /* cycle initiated */
        summ += delta;
        s += (t<<4) + E[0] ^ t+summ ^ (t>>5)+E[1];
        t += (s<<4) + E[2] ^ s+summ ^ (s>>5)+E[3];
    } /* end of the cycle*/
    E[0] = s ; E[1] = t ;
}

```

**Fig. 2** Tiny encrypting encode routine in C language

The performance of time of TEA on a workstation is so much remarkable. Because of using this algorithm, the difference of 1 bit in the plaintext would cause the difference of 32 bit in the coded text or ciphertext.

Figure 2 represents the TEA encode routine which is written using C programming language, where E[0], E[1], E[2], and E[3] are represented as the key values and d[0] and d[1] are the data values.

In Fig. 3, we represent the configuration of the TEA encryption method. The cleartext or plaintext and the key K are given as a input to the TEA encryption algorithm. According to the figure, the value of cleartext is  $C = (L[0], R[0])$  and the value of ciphertext or codedtext or secrettext is  $S = (L[64], R[64])$ . We have to divide the block of cleartext into two parts as  $L[0]$  and  $R[0]$ . The other half part is encrypted by the previous half part over 64 cycles of processing and after that these will combines to create the secrettext block.

- In every cycle, j has two insertion  $L[j-1]$  and  $R[j-1]$ , which is obtain from the prior round, furthermore, from the 128 bit overall E it also derived sub key  $E[j]$ .
- The sub keys  $E[j]$  have unlike values from each other and likewise from E.
- The constant value of delta equals to  $(\sqrt{5}-1)*231 = 2,654,435,769$  integer value, which is resulting from the golden number ratio, to be sure that the sub keys are dissimilar and the specific value of delta must not have cryptographic implication.
- In this algorithm, the cycle function is differs to some degree from a classical Feistel cipher structure. In this organization, as an alternative of Ex-OR the integer addition modulo  $2^{32}$  is used.

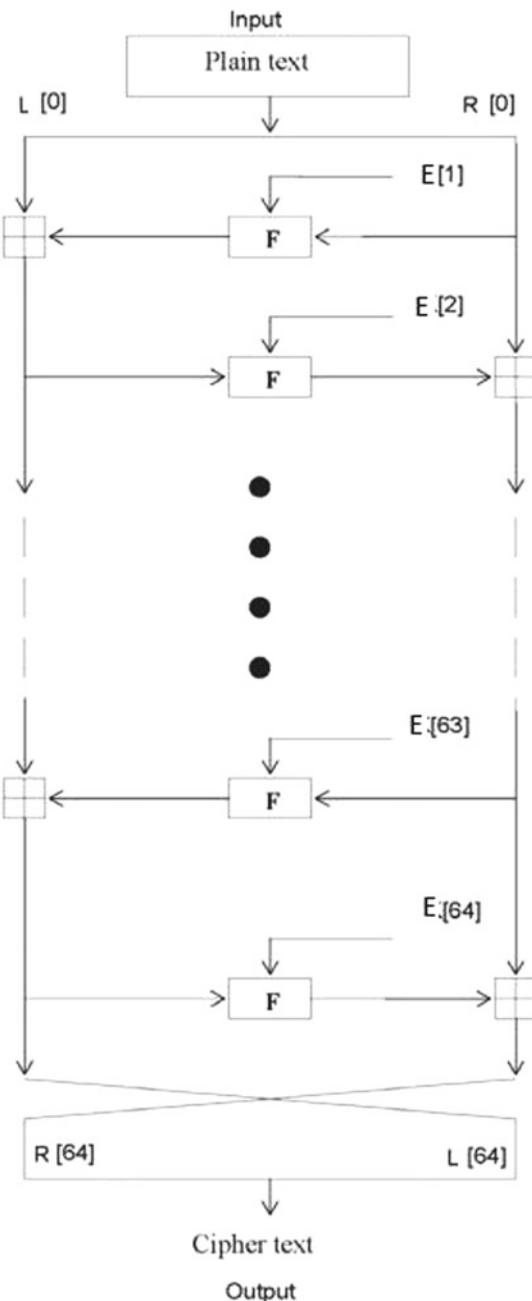
In Fig. 4, we represent the inside details of the jth rotation of TEA. The addition of key, bitwise EX-OR operation, and shift left and right operation are performed in the turn function F. The description of the output  $(L[j+1], R[j+1])$  of the jth round of TEA using its input as:

Following are the calculation of  $(Left[n], Right[n])$

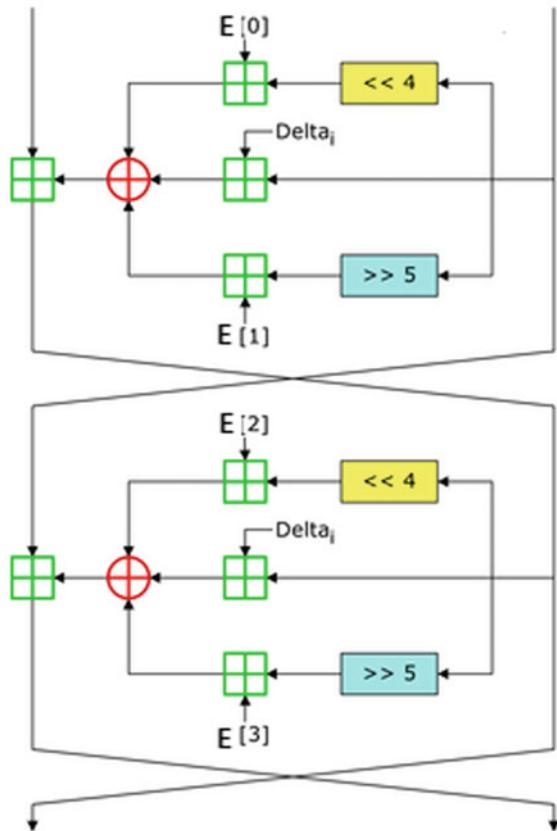
$$L[j+1] = L[j] F(R[j], E[0, 1], \text{delta}[j]),$$

$$R[j+1] = R[j] F(R[j+1], E[2, 3], \text{delta}[j]),$$

$$\text{delta}[j] = (j+1)/2 * \text{delta},$$



**Fig. 3** Tiny encryption structure



**Fig. 4** jth cycle of TEA

The cycle function,  $F$ , is defined as

$$F(M, K[p, q], \text{delta}[j]) = ((M << 4) + [p]) \oplus (M \text{ delta}[j]) \oplus ((N >> 5) + E[q]).$$

For each cycle, the cycle function  $F$  has the same structure but is parameterized by the cycle sub key  $E[j]$ . Here, we explain the key scheduling algorithm which is simple in nature; the 128-bit key value  $E$  is divided into four 32-bit blocks  $E = (E[j])$  where  $j = 0, 1, 2, 3$ ). Here, we used the keys  $E[0]$  and  $E[1]$  in the odd cycles and the keys  $E[2]$  and  $E[3]$  in even cycles.

In the decryption process (Fig. 5), we are using the same process of encryption method; in this routine, the process take input as the cipher text block, but in this method we will do use the sub keys  $k[j]$  in the reverse order.

In Fig. 6, we represent the decryption process of the TEA. In this, the intermediary value of the decryption procedure is same to the equivalent value of the encryption method with the two equal parts of the swapped value. Let's say, output of the  $j$ th encryption cycle is

$$EL[j] \parallel ER[j] \quad (\text{EL}[j] \text{ concatenated with ER}[j]).$$

```

void DEcode(long*d , long* E) {
    unsigned long n1= 32, summ, s = d[0], t = d[1],
    Delta = 0x9e3779b9 ;
    summ = delta<<5 ;
    /* start of round */
    while (n1-->0) {
        t = (s<<4)+e[2] ^ s+summ ^ (s>>5)+e[3] ;
        s= (t<<4)+e[0] ^ t+summ ^ (t>>5)+e[1] ;
        summ-= delta ; }
    /* end of round */
    d[0] = s ; d[1] = t ; }

```

**Fig. 5** Tiny decryption encode routine in C language

At that point, the matching input to the (64-j)th decryption cycle is now DR[j] || DL[j] (DR[j] concatenated with DL[j]).

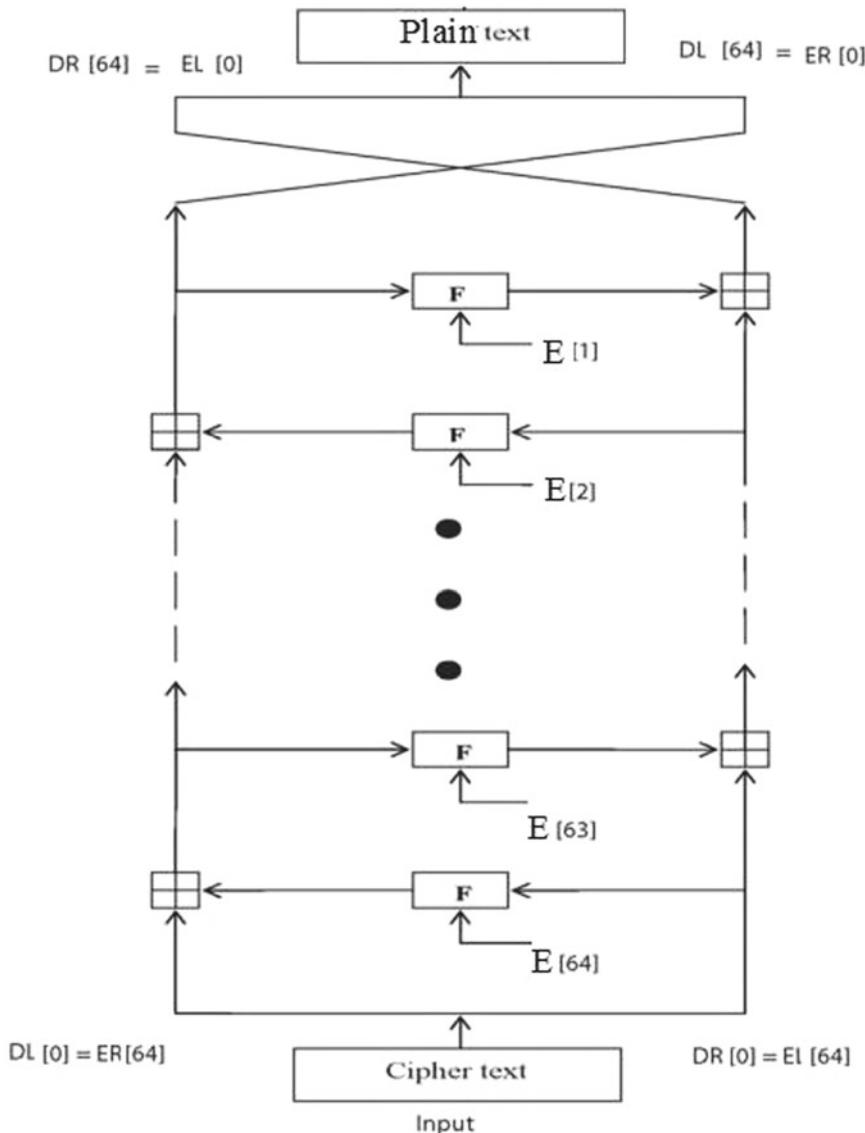
The two equal parts of the output will be interchanged after the last iteration of the encryption routine, so that the ciphertext is ER[64] || EL[64], the output which will get from this cycle is the final cipher text CT block. At this time, we shall use input to the decryption process as cipher text block. In this decryption method, the input to the first cycle is ER[64] || EL[64], equal to the 32-bit interchange of the output of the 64th cycle of the encryption method.

## 5.2 Working

The input data from the IoT device is given to the TEA algorithm. It is a light weight cryptographic algorithm which provides a secured data. This data is uploaded on the cloud server, and the original data can be retrieved from the cloud server by using inverse TEA algorithm.

## 6 Conclusion

The coming era, the necessary component of our everyday life will be Internet of Things. Using the IoT various sensors of the IoT devices are incessantly communicating with one another with energy constrained. The security of that communication must not be negotiated. Due to this security impetus, a lightweight algorithm TEA is suggested in this paper with a simple and integrated countermeasure against DoS in IoT. The security includes the procedures and controls that guarantee confidentiality, integrity, and availability of the information processed. However, there are some of the attack techniques of DoS such as attack tools, application-layer floods, degradation of services, and denial of services.



**Fig. 6** Tiny decryption structure

The implementation shows favorable results that make the algorithm an appropriate aspirant to be agreed in applications of IoT. As far as forthcoming scope is concern, we will be considering the complete performance assessment and crypto analysis of this procedure on various software and hardware platforms for known

probable attacks. The proposed algorithm will be compared against present security algorithms for security effectiveness and their effect on memory as well as computational usage.

## References

1. Grabovica M, Popić S, Pezer D, Knežević V (2016) Provided security measures of enabling technologies in Internet of Things (IoT): a survey. Zooming Innovation in Cosr Elect Intnl Conf
2. Navarro M, Davis TW, Villalba G, Li Y, Zhong X, Erratt N (2014) Towards long-term multi-hop WSN deployments for environmental monitoring: an experimental network evaluation. JSAN 3:297–330
3. Saha S, Matsumoto M (2007) A framework for disaster management system and WSN protocol for rescue operation. In: TENCON IEEE Conference, pp 1–4
4. Vaidehi V, Vardhini M, Yogeshwaran H, Inbasagar G, Bhargavi R, Hemalatha CS (2013) Agent based health monitoring of elderly people in indoor environments using wireless sensor networks. Procedia Comp Sci 19:64–71
5. Xu R, Huang X, Zhang J, Lu Y, Wu G (2015) Software defined intelligent building. Int J Inf Secur Privacy (IJISP) 9(3):84–99
6. Linlu L (2014) Comparative study on the development of internet of things policy in China and European Union Based. Sci Tech Info Dev Ec
7. Zhao K, Ge L (2013) A survey on the Internet of Things Security. In: ICCIS, pp 663–667
8. Alanazi S, Al-Muhtadi J, Derhab A, Saleem K (2015) On resilience of Wireless Mesh routing protocol against DoS attacks in IoT-based ambient assisted living applications. In: ICENAS
9. Alanazi S, Al-Muhtadi J, Derhab A, Saleem K (2015) On resilience of Wireless Mesh routing protocol against DoS attacks in IoT-based ambient assisted living applications. In: ICENAS
10. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comp Netwks 54: 2787–2805
11. Katagi M, Moriai S (2008) Lightweight cryptography for the internet of things. Sony Corp 7–10
12. Ebrahim M, Khan S, Mohani SSUH (2014) Peer-to-peer network simulators: an analytical review
13. Ebrahim M, Khan S, Khalid UB (2014) Symmetric algorithm survey: a comparative analysis. IJCA 61(20):0975–8887
14. Lim CH, Korkishko T (2005) mcryptona—a lightweight block cipher for security of low-cost RFID tags and sensors. Info Sec App Springer, pp 243–258
15. Lim CH (1998) Crypton: a new 128-bit block cipher. In: NIST AE Prop.
16. Engels D, Fan X, Gong G, Hu H, Smith EM (2009) Ultralight weight cryptography for low-cost RFID tags: Hummingbird algorithm and protocol. CACR Tech Rpts 29
17. Engels D, Saarinen MJO, Schweitzer P, Smith EM (2011) The hummingbird-2 lightweight authenticated encryption algorithm. RFID Sec. & Privacy. Springer, pp 19–31
18. Zhang K, Ding L, Guan J (2012) Cryptanalysis of hummingbird-2. IACR Crypt. ePrint Arc., vol 2012, p 207
19. Ganesan P, Venugopalan R, Peddabachagari P, Dean A, Mueller F, Sichitiu M (2003) Analyzing and modeling encryption overhead for sensor network nodes. In: Proceedings of the 2nd ICWSNA. ACM, pp 151–159
20. Schneier B (2007) Applied cryptography: protocols, algorithms, and source code in C. Wiley
21. Lai X (1992) On the design and security of block ciphers. Ph.D. diss., Diss. Techn. Wiss ETH Zürich, Nr. 9752, 1992. Ref.: Massey JL, Korref, uhlmann HB
22. <https://www.concise-courses.com/5-major-types-of-dos-attack/>

# Comparative Study of SVM and Naïve Bayes for Mangrove Detection Using Satellite Image



Anand Upadhyay, Santosh Singh, Nirbhay Singh, and Ajay Kumar Pal

**Abstract** Mangroves are a kind of plant which assumes an extremely fundamental job for security of our biological system. We presented the better approach for mangrove discovery by utilizing the help vector machine (SVM) and Naïve Bayes both are going under managed AI, and this calculation is utilized to group the image. The high-goals satellite information from Google earth is procured from an alternate locale of Mumbai, Maharashtra district, for recognition of mangroves. This exploration paper utilized two unique calculations, for example, Naïve Bayes classifier and Support Vector Machine for the discovery of perusing highlights from satellite images, and there are two calculations which are actualized utilizing the Matlab recreation tool stash. Support Vector Machine and Naïve Bayes are a directed grouping strategy applied on satellite image. In the wake of applying the calculations on the picture satellite, the precision of classifiers is determined utilizing perplexity grid and kappa coefficient. The execution of both methods of Support vector machine and Naïve Bayes generate the detected area of mangrove in Mumbai, Maharashtra region. Exactness of Naïve Bayes saw as 99% with kappa value 0.9831, and the precision of help vector machine saw as 97% with a kappa estimation of 0.9631. The precision figuring utilizing disarray lattice and kappa coefficient shows that the Naïve Bayes classifiers is superior to help vector machine for the discovery of mangroves utilizing satellite picture.

**Keywords** Machine learning · MATLAB · Remote sensing · Google earth · Satellite image

---

A. Upadhyay · S. Singh · N. Singh (✉) · A. K. Pal  
Thakur College of Science & Commerce, Kandivali (E), Mumbai 400101, India  
e-mail: [nirbhaysingh69682@gmail.com](mailto:nirbhaysingh69682@gmail.com)

A. Upadhyay  
e-mail: [anandhari6@gmail.com](mailto:anandhari6@gmail.com)

S. Singh  
e-mail: [sksingh14@gmail.com](mailto:sksingh14@gmail.com)

A. K. Pal  
e-mail: [pal72121@gmail.com](mailto:pal72121@gmail.com)

## 1 Introduction

In science, mangroves are a kind of plant which assumes an extremely fundamental job for security of our biological system. We presented the better approach for mangrove discovery by utilizing the help vector machine (SVM) and Naïve Bayes both are going under managed AI, and this calculation is utilized to group the image. The mangroves are a sort of plant which can be experienced in childhood in water-front area. The high-goals satellite information from Google Earth is used to procure from an alternate locale of Mumbai, Maharashtra district, for mangrove recognition. This exploration paper utilized two unique calculations, for example, Naïve Bayes classifiers and Support Vector Machine for the discovery of perusing highlights from satellite images, and the two calculations are actualized utilizing the Matlab recreation tool stash. Support Vector Machine and Naïve Bayes is a directed grouping strategy applied on satellite image. In the wake of applying the calculations on the picture satellite, the precision of classifiers is determined utilizing perplexity grid and kappa coefficient. The execution of both methods of Support vector machine and Naïve Bayes generate the detected area of mangrove in Mumbai, Maharashtra region. Exactness of naïve Bayes saw as 99% with kappa value 0.9831, and the precision of help vector machine saw as 97% with a kappa estimation of 0.9631. The precision figuring utilizing disarray lattice and kappa coefficient shows that the naïve Bayes classifier is superior to help vector machine for the discovery of mangroves utilizing satellite picture (Fig. 1).



**Fig. 1** Mangroves and water

## 2 Literature Review

Mangrove woodlands assume a significant job in giving biological and financial administrations to human culture. The future pattern of mangrove woods changes was anticipated by a Markov tie model to help basic leadership for seaside the board [1]. Mangrove timberlands give significant biological system products and ventures for human culture. Broad seaside advancement in many creating nations has changed over mangrove timberlands to other land utilizes regardless of their biological system administration coefficients; along these lines, the environment condition of mangrove backwoods is basic for authorities to assess feasible waterfront the executives techniques [2]. In a less reported part, mangrove trees moreover direct flooding, disintegration, siltation, and tidal surges, subsequently giving environments for people and incalculable other living beings. About two-thirds of all sorts of commercial angles too depend on mangroves for their development, generation, and survival [3]. Mangroves of Maharashtra are beneath incredible risk of human encroachment. The Raigad locale is near to the Mumbai, and the exercises in Mumbai have the reflection over the Raigad mangrove within the show work endeavor is made to dissect the zone occupied by the mangroves within the coastal area of Raigad area and the changes within the mangrove living space over the period of time utilizing Google SoilMaster pictures and ground perceptions [4]. Nowadays due to the increase in coastal society, industrial development and urbanization, mangrove forests are decreasing day by day and it causes natural disasters such as hurricanes and tsunamis. For conservation, proper management and restoration measures of the coastal region, updated information about mangrove ecosystems is essential for us [5]. Mangrove woods are among the most gainful environments on Earth and fundamentally develop at tropical and subtropical scopes. They give numerous significant biological and cultural capacities [6]. In India, mangroves are found on the east and west shorelines of the territory and on the Islands of Andaman and Nicobar and Lakshadweep. Indian mangroves address 3.3% of overall mangroves and about 56% of overall mangrove species [7]. Mangrove bogs and backwoods are a fundamental interface of the beach front zone that give different natural and monetary administrations adding to Waterfront assurance and carbon credits. Remote-detecting systems are routinely used to give spatial-worldly data on mangrove biological system conveyance, species distinguishing proof, wellbeing status, and populace [8]. To order mangrove species, remote-detecting advances furnish a superior path with high spatial goals picture.

The spatial structure is normally seen as viable integral data for arrangement. Be that as it may, it is as yet a test to configuration handmade highlights for mangrove species due to their non-structure surface [9]. Mangroves spread under 0.1% of Earth's surface, store a lot of carbon for each unit zone, and however are undermined by worldwide natural change [10].

### 3 Characteristics of Data and Field Study

The any kind of research completely depends upon the data and types of data used for processing and finding the results. The data plays very effective and major role in research. Here, the data from satellite image is stored from Google Earth imaginary platform. The Google Earth imaginary delivers the data of high resolution. When any types of satellite image have high resolution of data then in such types of images, it is very easy to differentiate the features from each other and it is also easy to collect the features for detection, recognition, and classification of any resources or objects.

### 4 Methodology

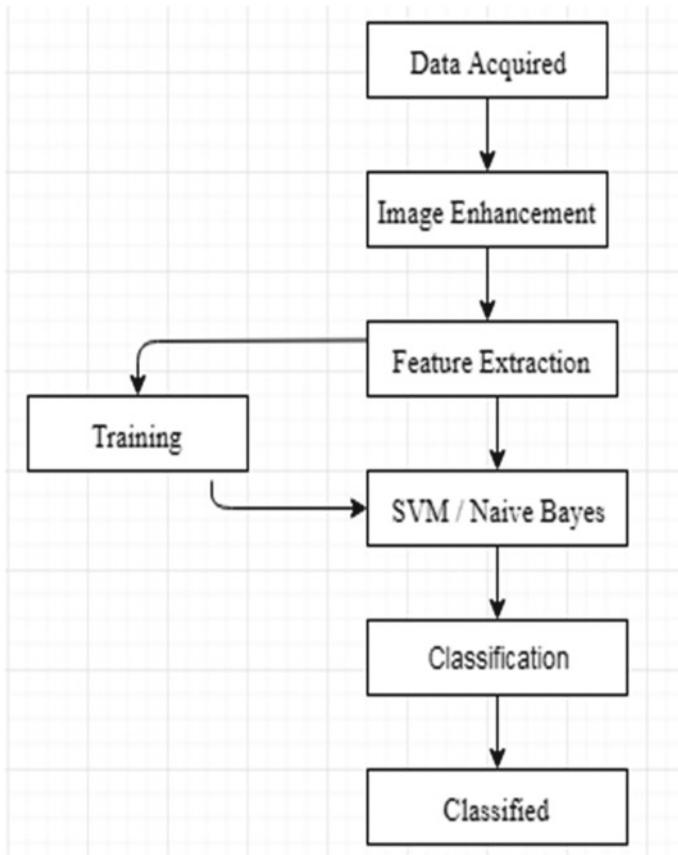
The proposed algorithms are implemented using MATLAB R2010 simulation and mathematical toolkit. The MATLAB is very famous and robust mathematical simulation toolkit which is specifically designed for processing and implementation of mathematical model. The SVM and Naïve Bayes both are executed using the simulation toolbox from MATLAB. Here, the data which is used as Google Earth high-resolution satellite image for classification of mangrove region. The working model of the proposed methodology is presented with the help of presented flow chart (Fig. 2).

#### A. Support Vector Machine (SVM)

SVM is called as Support Vector Machine which belongs to very powerful classification algorithms. The SVM widely used and proposed to use for the dataset or classification problem where the data is huge in amount. The SVM generated hyperplane with the help of datasets and try to fit the data with respect to proposed or generated hyperplane. The SVM gives the binary classification problem-solving techniques but for multiclass classification the one versus all methodology is used for such types of problem. Here, in the proposed concept our problem needs binary classification; therefore, there is no need to apply one versus all concept. The MATLAB R2010 provides the method for implementation of SVM which required the training datasets along with its specified classes.

#### B. Naïve Bayes Classifier

Naïve Bayes classification techniques are also called as probabilistic stratification techniques. Naïve Bayes classification technique works on the principle of Bayesian probability techniques. The Bayesian algorithms provide the relation among different attributes and with the help of probability value the different associations are created, i.e., high association, moderate association, or low association. Here, in the MATLAB R2010 proves the simulation method for implementation of Naïve Bayes classification techniques. The method required training dataset and specified classes for different attributes.



**Fig. 2** Hierarchy of classification

## Classification

In light of the preparation model, the framework will play out the arrangement of the picture in two gathering mangroves and non-mangroves. The framework will show the mangrove parcel as red shading, and remaining part will be same. In view of this order, disarray lattice, kappa coefficient, and exactness are obtained.

## 5 Results

SVM prepared along with example information. In the wake of preparing, testing is performed utilizing testing dataset. In the wake of preparing, exactness and kappa coefficient are gotten by perplexity framework and characterization report is created (Table 1).

**Table 1** Error matrix for SVM classifier

Classes	Mangroves	Non-mangroves	Total	User accuracy (%)
Mangroves	145	8	153	94.77
Non-mangroves	1	174	175	99.42
Total	146	182	328	
Producer accuracy (%)	99.31	95.60		

$$= \mathbf{97.25\%}$$

$$K = \frac{N \sum_{i=1}^r Xii - \sum_{i=1}^r (Xi + *X + i)}{N2 - \sum_{i=1}^r (Xi + *X + i)}$$

$$\mathbf{k = 0.9631}$$

The above calculation shows the performance and efficiency of SVM classifier using the kappa coefficient and error matrix. The results show accuracy of 97.25% which is good and expected from Support Vector Machine classification technique (Figs. 3, 4 and 5; Table 2).

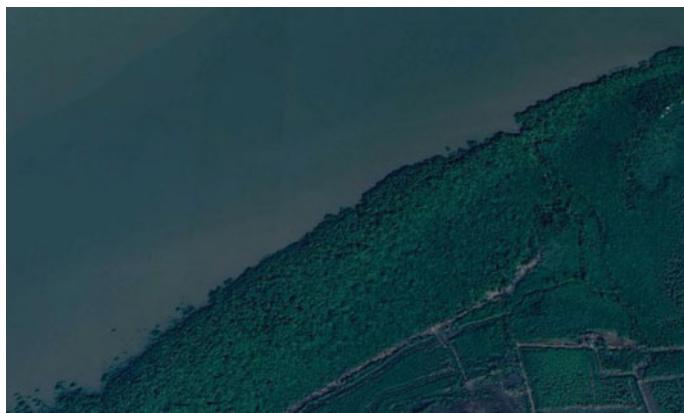
$$\text{Accuracy} = (386/389) * 100$$

$$= \mathbf{99.22\%}$$

$$K = \frac{N \sum_{i=1}^r xii - \sum_{i=1}^r (Xi + *X + i)}{N2 - \sum_{i=1}^r (Xi + *X + i)}$$

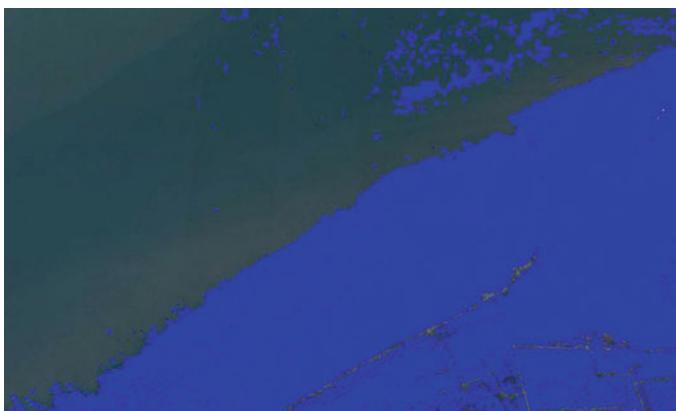
$$\mathbf{K = 0.9831 (Good)}$$

The Naive Bayes classifier's error matrix calculated the accuracy of 99.22% for classification of Google Earth which is high than the accuracy of Support Vector

**Fig. 3** False color image before classification



**Fig. 4** Naïve bayes classification



**Fig. 5** Support vector machine classification

**Table 2** Error matrix for naïve bayes classifier

Classes	Mangroves	Non-mangroves	Total	User accuracy (%)
Mangroves	178	2	180	98.88
Non-mangroves	120	8	209	99.52
Total	179	210	389	
Producer accuracy (%)	99.44	99.004		

Machine. In the case of proposed classification of high-resolution satellite images from Google Earth and there performance of Naïve Bayes is good compared to SVM.

## 6 Conclusion

In this paper, we effectively accomplished the mangrove's location with assistance of the Support Vector Machine and Naïve Bayes classifier calculation which separated the mangroves from the satellite picture and give the distinctive shading to mangroves and it is obvious to the end client. In the way of implementing, both calculation characterizations are applied which can be seen as the diverse exactness on each of the calculation. Naïve Bayes calculation gives the great exactness than the Support Vector Machine algorithm; from this examination, it is presumed that Naïve Bayes calculation is giving higher precision than the Support Vector Machine.

## 7 Future Enhancement

The above investigation is accomplished by utilizing Naive Bayes and Support Vector Machine for Mangrove recognition from satellite images. For the future examination, other arrangement calculations such as KNN, artificial neural system (ANN), and so forth can be utilized for mangrove location utilizing satellite picture and for their near investigation.

## References

1. Chen C-F et al (2013) Multi-decadal mangrove forest change detection and prediction in Honduras, Central America, with Landsat imagery and a Markov chain model. *Remote Sens* 5(12):6408–6426
2. Son N-T et al (2014) Mangrove mapping and change detection in Ca Mau Peninsula, Vietnam, using Landsat data and object-based image analysis. *IEEE J Selected Top Appl Earth Obs Remote Sens* 8(2) :503–510
3. Gevana D et al (2019) Land use characterization and change detection of a small mangrove area in Banacon Island, Bohol, Philippines using a maximum likelihood classification method. *Forest Sci Technol* 11(4):197–205
4. Telave AB, Ghodake SD, Pawar GP (2017) Studies on area assessment under mangroves of Raigad District, Maharashtra, India. *Indian Forester* 143(3):207–212
5. Ghorai D, Mahapatra M, Paul AK (2019) Application of remote sensing and GIS techniques for decadal change detection of mangroves along Tamil Nadu Coast, India. *J Remote Sens & GIS* 7(1): 42–53
6. Ma C et al (2019) Change detection of mangrove forests in coastal Guangdong during the past three decades based on remote sensing data. *Remote Sens* 11(8):921
7. Ragavan P et al (2019) Current understanding of the mangrove forests of India. *Research Developments in Saline Agriculture*. Springer, Singapore, pp 257–304

8. Saravanan S et al (2019) Utility of landsat data for assessing mangrove degradation in Muthupet Lagoon, South India. *Coastal Zone Management*. Elsevier, pp 471–484
9. Wan L et al (2019) A small-patched convolutional neural network for mangrove mapping at species level using high-resolution remote-sensing image. *Annals of GIS* 25(1):45–55
10. Vázquez-Lule A et al (2019) Greenness trends and carbon stocks of mangroves across Mexico. *Environ Res Lett* 14(7):075010 (2019)

# Identification and Assessment of Black Sigatoka Disease in Banana Leaf



Anand Upadhyay, Neha Maria Oommen, and Siddhi Mahadik

**Abstract** Detecting a disease in plants is one of the challenging works. Identifying the disease through naked eyes is difficult. India is famous for agriculture. There were no modern techniques used in machine learning to find disease in banana leaf. Diseases like bacterial wilt and Black Sigatoka in banana leaf cause massive loss to the farmers. With the help of image processing technique and support vector machine algorithm, we can detect the disease called Black Sigatoka in banana leaf. Since this technique is cost effective, it is helpful for the farmers and one can easily detect the disease.

**Keywords** Banana leaf · Black Sigatoka · Disease detection · Image processing technique · Support vector machine

## 1 Introduction

Plants have become an important source of energy. The world is moving more toward technology dependent era. Every day we keep hearing woes of farmers that even after using costly fertilizers the leaves were eaten away by various diseases. Detection of leaf diseases plays an important role in today's world. It helps to minimize the work of farmers. Taking care of plants and detecting diseases in the early stage are crucial. Mycosphaerella fijiensis causes Black Sigatoka disease in leaves of various plants. It is also called as black leaf streak. Small lesions are present parallel to the veins of banana leaf. It can be controlled by costly fungicide spray [1]. In early days, the

---

A. Upadhyay · N. M. Oommen · S. Mahadik (✉)

Department of Information Technology, Thakur College of Science and Commerce,  
Kandivali (E), Mumbai 400101, India

e-mail: [siddhimahadik.sm@gmail.com](mailto:siddhimahadik.sm@gmail.com)

A. Upadhyay

e-mail: [anandhari6@gmail.com](mailto:anandhari6@gmail.com)

N. M. Oommen

e-mail: [nehamaria23@gmail.com](mailto:nehamaria23@gmail.com)

practiced or any skillful person manually analyzed the disease by working for hours and a lot of time was wasted. Image processing is one of the best techniques used to detect disease faster in plants. This technique is very easy and less expensive [2]. It will identify the affected area of disease, and the texture, shape, and colors are also determined. Machine vision, image processing, and computer vision are the advanced computer technologies, and it provides feasible support to the growers. This research paper outlines possible plant diseases that affect the leaves of banana plant and it completely checks different image processing techniques used for classification identification. Due to a vast number of plants getting affected by such diseases, it takes a great toll on the quality, quantity, and productivity of the vegetation which indirectly affects the economy and also health of mankind. For early detection of such diseases, it requires intensive monitoring procedures and observation by experts or farmers making it more expensive and time-consuming. There are numerous methods for detecting plant disorders. Most of the diseases are asymptomatic and only appear when it becomes worse. Usually, it is detected by means of advanced measures like use of microscopes. Commonly used methods are remote sensing method that captures multi- and hyperspectral photos. To achieve this, one can use digital image processing instruments [3].

## 2 Literature Review

Generally, it has been seen from the literature that NN has been applied for the identification of plant diseases, whereas the learning ability of SVM also contributes for the same purpose. However, the results obtained from real-life images are very encouraging. There are various implementation methods such as image processing, visual analysis, and optical sensor. The system can be developed by using these three methods to detect the disease earlier and can overcome the disadvantages and challenges [4]. By comparing the methods, in visual analysis it does not give the correct output, and in optical sensor, it is difficult to implement and expensive too. So, the image processing technique is the best way to build robust, simple, and accurate disease detection system. The database collection is the most challenging work while using image processing. For collecting data, it is essential to collect details of the crop and the disease from which it is suffering. The different types of leaf disorders can be fungal, viral, bacterial, or due to insects. This proposed paper will identify a single leaf disease [4].

Sannakki and Rajpurohit [5] proposed the work based on back propagation neural network and k-means segmentation for the detection of pomegranate leaf diseases. RGB image is transformed to  $L * a * b$  space and color and texture features are extracted from the image. The drawback of this technique is that it is applicable only for limited crops. Vipinadas and Thamizharasi [5] proposed a method which uses image processing which is similar to our paper. They have also used ANFIS classifier but instead of that we are using SVM method to detect Black Sigatoka disease in banana leaf as it gives better results.

### 3 SVM Algorithm

SVM stands for support vector machine. Support vector machine used for both regression and classification challenges. It is a supervised machine learning algorithm. Mostly, used in classification problems. It is best for separating two classes. It performs well with a less amount of data and it is fast too. This algorithm is used to perform image classification. Most of the time image processing is done before SVM. Support vectors are the points we find closest to the line from both the classes in svm algorithm [3]. If there are errors in the sample, it is robust. It gives high accuracy for prediction. Unlike neural networks, the complexity of computing SVM does not depend on the dimensions of the input space

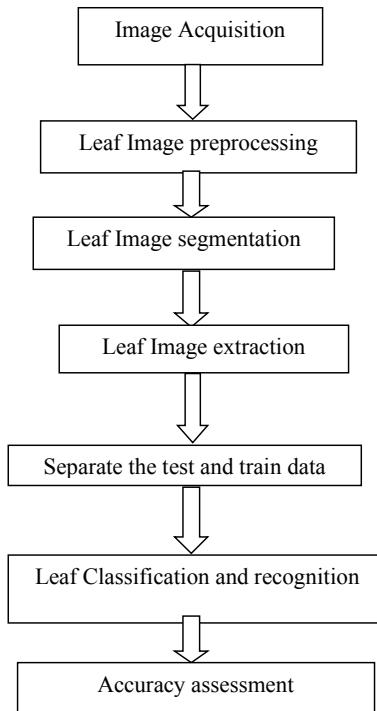
- Load the image.
- Preprocess the acquired image.
- Apply SVM.
- Extract the features from it.
- Classifier will classify the leaf.

### 4 Dataset

Firstly, the dataset folder is created which consists of various banana leaf images required. The banana plant's leaf image of size  $256 \times 256$  pixel is taken as an input. Two main files were generated as Training file and the Testing file. Each row represent the class that is the affected area of the particular row of data as class 1 and the unaffected area of the particular row of data as class 2. Then, a software program will be written in matlabR2016a that would take in .mat files representing the training and testing data, train the classifier using the train files and then use the test file to perform the classification task on the test data. It will load all the data files (training and testing data files) and make changes to the data according to the proposed model chosen. The images should have keen texture details of leaf for this method to produce better result.

### 5 Methodology

Different types of leaf images are taken in, and then, those are used to recognize the affected area in leafs. Then, various types of classifications techniques are applied on them, to classify those unaffected and affected banana leaf images. We are using MatlabR2016a for this proposed topic. Image processing techniques are utilized for the purpose of creating automated system in identification of diseases depending on various limitations. Green colored leaves indicate the health of the banana plant. The diseased leaf appears in yellow, black, or brown color. The change in leaf color is an



**Fig. 1** Proposed system for plant leaf classification

important parameter which determines whether the leaf is diseased or not. The steps used for plant leaf disease detection are as follows (Fig. 1).

### 5.1 *Image Acquisition*

In disease analysis, capturing the image is the first step. The images of the leaves are generated either whole or as a fraction depending on the analysis. These images will be obtained usually in RGB color format. The captured images are then converted to RGB model [6] (Fig. 2).

### 5.2 *Image Pre-processing*

It is a manipulation process which is performed on an image to provide information about the infected part in a leaf. It provides proper information of an image with proper visual understanding. It changes dynamic range of specific features for the



**Fig. 2** RGB2GrayScale

purpose of localization but it does not change default information in the leaf image. Image resizing is the common pre-processing technique used in detection of banana leaf disease.

### **5.3 Image Segmentation**

Different parts of images are separated to the region of interest. This process is to divide the image into regions.

This technique can be contextual or non-contextual. Accurate dividing of an image is a challenging problem.

### **5.4 Feature Extraction**

It is the process of dimensionality reduction. It captures the main part of an image. Initial set of raw data is reduced to groups for further processing. This process is useful when we need to reduce number of resources without losing any information. It is performed after the pre-processing technique. After the feature extraction classification is done, the classifier will assign a label to the image indicating which class it belongs. Here, support vector machine is used to classify the image [3].

## 6 Result and Analysis

Here, we have used matlabR2016a as our computer language. The performance of a classifier has been evaluated using confusion matrix (Figs. 3, 4; Table 1).

$$\begin{aligned} K &= \frac{N \sum_{i=1}^r xii - \sum_{i=1}^r (Xi + *X + i)}{N2 - \sum_{i=1}^r (Xi + *X + i)} \\ &= 0.8135(\text{Good}) \end{aligned}$$

Above is the confusion matrix and kappa coefficient (K) for the dataset which is tested by support vector machine algorithm.



**Fig. 3** Original image



**Fig. 4** Affected area detected

**Table 1** Confusion matrix

Class	Affected	Non-affected	Total	Procedure accuracy (%)
Affected	284	1	285	99.64
Non-affected	28	486	514	94.55
Total	312	487	799	
User accuracy (%)	91.21	99.74		96.37

$$\begin{aligned} \text{Accuracy} &= (770/799)*100 \\ &= 96.37 \end{aligned}$$

**Accuracy = 96%**

## 7 Conclusion

From the results shown above, we got the accuracy 96%. The images are classified using support vector machine algorithm. Compare to other algorithm the support vector machine (SVM) classifier gives more precise result. The enhanced system uses machine learning approaches to ensure that only relevant features are extracted. Therefore, we have presented a reliable system which would definitely help the farmers in the early stage and this approach will also increase the production of crops and improve the quality of the food. Also by computing severity and amount of disease present on the crop, only necessary and sufficient amount of pesticides can be used making agriculture production system economically efficient. So there is a scope of improvement [7].

## 8 Future Enhancement

The existing technology takes still image as the input while the future work is to create video input. To ensure greater efficiency and effectiveness, we need a grading system which enables us to prioritize the disease according to their intensity. We can also improve using better segmentation, feature extraction to increase the recognition rate at the final stage, thus obtaining more vital system.

**Acknowledgements** The authors particularly wish to acknowledge all the teachers for their support, encouragement, and invaluable guidance in preparation of this research. Thanks to the kaggle.com website for the dataset.

## References

1. Prabukumar M, Balamurali J (2014) Image processing and pattern classification technique in a machine vision system that identifies and classifies the plant diseases based on the visual symptoms. *Int. J. Adv Res Comput Sci*
2. Al Hiary H, Bani Ahmad S Reyalat M (2014) Fast and accurate detection and classification of plant diseases. *Int J Comput Appl*
3. Namrata K (2017) Leaf based disease detection using “GLCM and SVM”. *Int J Sci Eng Technol*
4. Ijsea.com (2019) [Online] Available:<https://www.ijsea.com/archive/vol.7/issue8/IJSEA0708/003>
5. Thamizharasi A (2016) Detection and grading of diseases in banana leaves using machine learning 7(7)
6. Surya P, Kumar S (2013) Assessment of banana fruit maturity by image processing technique. *J. Food Sci. Technol*
7. Mainkar P, Ghorpade S, Adawadkar M (2015). Plant leaf disease detection and classification using image processing techniques. *Int J Innovative Emerg Res Eng*

# Water Resource Detection Using High Resolution Satellite Image and GRNN



Anand Upadhyay, Manisha Pandey, and Ajay Kumar Pandey

**Abstract** Water is the most important for human body, environment, and transportation and so on. One of the interesting areas of research is the use of satellite image. We can use the various techniques and compare them to check the result to each other. Survey the very dry area for water and conventional techniques used the satellite image. The objective of the paper is water resource detection using satellite image. Satellite image provides the data and information about any earth surface or object without making physical contact with it. It will help to come up with the best idea or technique that can be used for our research. We have used GRNN algorithm to detect the water resources that are available on the surface of earth, and we found 97.10% accuracy.

**Keywords** Satellite image · GRNN · Water resource · Earth · Remote sensing

## 1 Introduction

The every living organism needs water, food and air which are there compulsory need. Therefore, water is vital resources available on the earth surface. Specially, the water plays important role in human's day to day life. There are lots of industry and organization which cannot run without water. But simultaneously, these all industries also releasing there bad, polluted and chemical affected water in natural water bodies and spoiling and damaging the natural water bodies and resources and its impact is in terms of hazardous disease, health problem and damage of water ecosystem. So, it is very important to monitor the water bodies and for that we have to first recognize

---

A. Upadhyay (✉) · M. Pandey · A. K. Pandey  
Thakur College of Science and Commerce, Kandivali (E), Mumbai 400101, India  
e-mail: [anandhari6@gmail.com](mailto:anandhari6@gmail.com)

M. Pandey  
e-mail: [manishapandey.05050@gmail.com](mailto:manishapandey.05050@gmail.com)

A. K. Pandey  
e-mail: [ajaypanday678@gmail.com](mailto:ajaypanday678@gmail.com)



**Fig. 1** Water and other information

the water bodies with the help of different techniques. The remote sensing is one of the best and very appropriate techniques which can help to solve this problem because there is no human intervention involved. Here, the Google earth based high resolution satellite image is used to recognize and classify the water bodies on earth surface. The general regression neural network is proposed model which is suggested here for classification of water bodies. The general regression neural network is one type of artificial neural network which is used for classification problem. Such types of model for recognition of different resources available on earth surface help to develop the sustainable model for future development because day by day the human population is growing and is very difficult to control anyone's need and desires. Every human being needs shelter, water, food and even luxuries life too; therefore, we need such types of model and mechanism to monitor different types of natural resources and water is one of them. The proposed techniques are implemented in different steps which was started from collection Google Earth satellite image, field visit and collection of training datasets, implementation of algorithm and last but not least classification and testing of datasets. The result of proposed technique was calculated by confusion matrix and Kappa coefficient (Fig. 1).

## 2 Literature Review

The aim of paper was to determine the accuracy of using digital image processing to map river water bodies with Landsat-5 satellites image. Landsat-5 image data is used to detect water source. Density dividing of the single mid-infrared band 4 proved as successful as multispectral classification achieving an overall accuracy of the output for that algorithm [1]. Water erosion can harm the agriculture and water quality, collected the past 30 years' data and find the changes in the water quality [2]. Another strategy depends on principal components of multi-fleeting NDWI was given and assessed for earth water change discovery. Results show an extreme diminishing pattern in Lake Uremia surface zone in the period 2001–2014,

particularly somewhere in the range of 2011 and 2014 when the lake lost around 33% of its surface region contrasted with the year 2001 [3]. Point by point wetland maps can be refreshed utilizing satellite symbolism. Given the spatial goals of satellite remote detecting frameworks, fully characterization, subpixel order, spectral mixture examination and blends estimation may give progressively definite information on wetlands. A layered, half breed or principle based methodology may give better results than progressively customary techniques. The mix of radar and optical information provide the most guarantees for improving wetland order [4]. This is regularly allowed to as hyperspectral symbolism. These audit subtleties the contrasts among multispectral and hyperspectral information, spatial and goals and spotlights on the utilization of hyperspectral symbolism in water asset pond and, specifically, the grouping and mapping of land uses and vegetation [5]. Results got in this research ought to hold any importance with the global group of spectators of environmental management in that they feature the intelligent idea of human exercises and the earth and the off-site effect of these exercises on the earth [6]. A similar evaluation of the water assets that are accessible in a neighbourhood district territory or a waterway is important for discovering answers for water-related issues concerning both the amount and nature of the water assets on the earth surface [7]. This technique can take fuzzy properties of upper qualities for grade benchmarks into a full record and abstain from deciding the vulnerability coefficient I inset pair examination [8]. Address such sort of difficulty; a suite of demonstrating methods to aid the assessment of characteristic and impacted waterway streams at ungauged locales has been created [9]. Next to each other, correlation apparatuses for various assessment measurements, including ROC bends, time to recognize, and bogus, alert rates which will discover the water assets [10].

### 3 Data and Field Study

Data has a vital role any kind of research and findings. The any kind of research completely depends upon accurate and précis datasets. The remote sensing provides an unbiased data without actual visit to that place and without any human involvement and intervention. In the proposed research work, the remotely detected information is gathered from Google Earth. Google Earth provides exact and cutting-edge dates. Google Earth gives the information wherever which is as of late caught a couple of days prior. The information which is gathered from Google Earth is high-goal satellite picture of the water. More than 96% water is saline water in the seas which have no use. The freshwater assets, for example, water moving into streams, waterways, lakes, and groundwater, the water they need each day to live. Water on the outside of the Earth is anything but difficult to imagine, and your perspective on the water cycle may be that precipitation tops off the waterways and lakes. The Google earth gives false colour image.

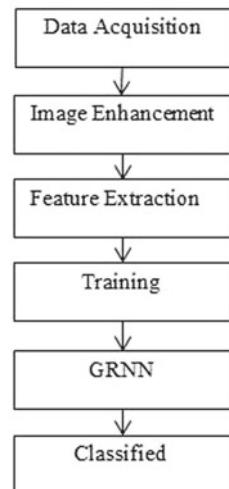
## 4 Methodology

The suggested algorithm is implemented using MATLAB R2010 simulation toolbox. The MATLAB provides the simulation toolbox with the function and features to implement the general regression neural network. Here, Google Earth datasets are used for processing which is coloured image consists of information in the form of colour bands. The following flow chart presents the implementation of entire work flow (Fig. 2).

### 4.1 Feature Extraction

The Google Earth gives the data in the form of colour image which consist of numerous information in the form of red, green and blue band. The first and primary step of GRNN-based classification is feature extraction and preparation of training and testing datasets. Here, pixel colour-based features are collected based on field study and help of Google map for different objects available on earth surface verses and water bodies. The collected data are divided into two different file for training and testing purpose.

**Fig. 2** Flow of methodology



## 4.2 GRNN Algorithm

GRNN is known as general regression neural network. It is one type of artificial neural network which design the parallel architecture for implementation neural network. The GRNN neural network implements the nonparametric regression for classification of data. The fundamental principle behind GRNN is function approximation and function estimation. The GRNN neural network represents associate degree improved technique.

### Algorithm

It predicts the output of a given input data. As per the basic principle of neural network, it needs a training data to train itself.

$$Y(x) = \frac{\sum_{k=1}^N y_k k(x, x_k)}{\sum_{k=1}^N K(x, x_k)}$$

where

- $Y(x)$  is the prediction value of input  $x$ .
- $y_k$  is the activation weight for the pattern layer neuron at  $k$ .
- $k(x, x_k)$  is the radial basis function kernel.

## 5 Experimental Results

GRNN prepared with the training data which we found from the classified water resource image. After performing the calculation on the image, we calculated the accuracy and Kappa value which is obtained by confusion matrix and classification report is generated in Table 1.

After calculating the accuracy of GRNN, we found that GRNN provides more accuracy than the KNN and decision tree algorithm which is used by the other researcher (Fig. 3).

**Table 1** Confusion matrix of GRNN algorithm

Classes	Water	Non-water	Total	User accuracy (%)
Water	278	20	298	97.10
Others	0	393	393	100
Total	278	413	691	
Produce accuracy (%)	100	95.15		

$$\text{Accuracy} = (671/691)*100 \\ = 97.10\%$$

**Fig. 3** Image after applying GRNN algorithm



## 6 Conclusion

Water is very important resource available on earth surface. It is life and need of every living body on the earth. In this research paper to achieve the defined goal or hypothesis for classification and recognition of water bodies, the Google Earth satellite image is used. With the help of study, Google map and field work, the training datasets are prepared for training of the general regression neural network. After design and training general regression neural network, it is tested using the testing datasets and it shown the accuracy of 97.10%. Therefore, from the proposed implemented method, the results show that the general regression neural network is very good for recognition of water bodies using Google Earth high resolution satellite image.

## 7 Future Scope

The suggested method used general regression neural network for classification of high resolution Google Earth satellite image for recognition of water bodies. In future, suggested algorithm can be tested on other types of satellite image, i.e. LISS-III, AWIFS, LISS-IV, etc. and there comparison can be performed for better understand of satellite image. This method can also tested using fusion of different satellite image by choosing more feature from satellite datasets. The results of general regression neural network are also compared with other neural network of neural network family.

## References

1. Frazier PS, Page KJ (2000) Water body detection and delineation with landsat TM data. *Photogram Eng Remote Sens* 66(12):1461–1468
2. Vrieling A (2006) Satellite remote sensing for water erosion assessment: a review. *CATENA* 65(1):2–18
3. Rokni, K., et al. (2014) Water feature extraction and change detection using multitemporal Landsat imagery. *Remote sensing* 6(5), 4173–4189
4. Ozesmi SL, Bauer ME (2002) Satellite remote sensing of wetlands. *Wetlands Ecol Manage* 10(5):381–402
5. Govender M, Chetty K, Bulcock H (2007) A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa* 33(2)
6. Liu Y et al (2005) Land use/cover changes, the environment and water resources in Northeast China. *Environ Manage* 36(5):691–701
7. Xu C-Y, Singh VP (2004) Review on regional water resources assessment models under stationary and changing climate. *Water Resour Manage* 18(6):591–612
8. Wang WS et al (2009) A new approach to water resources system assessment—set pair analysis method. *Sci China Ser E: Technol Sci* 52(10):3017–3023
9. Young AR, Grew R, Holmes MGR (2003) Low flows 2000: a national water resources assessment and decision support tool. *Water Sci Technol* 48(10):119–126
10. Hart D et al (2007) CANARY: a water quality event detection algorithm development tool. In: World environmental and water resources congress 2007: Restoring Our Natural Habitat

# Retinopathy Detection Using Probabilistic Neural Network



Anand Upadhyay, Parth Kantelia, and Rohan Parmar

**Abstract** We propose a diabetic retinopathy (DR) analysis algorithm based on probabilistic neural network (PNN). This algorithm is used to recognize the pattern problem. By this algorithm, we can help in diagnosis of a diabetic patient regarding their damage to the back of retina (eye) occurred in tissue of blood vessels using probabilistic neural network. PNN is also known as feed forward neural network. This algorithm has been tested on a small image database and compared with the performance of a human eye. Confusion matrix and kappa coefficient are used to find the accuracy rate of the diabetic eye.

**Keywords** Diabetic retinopathy · Probabilistic neural network · Matlab · Confusion matrix · Kappa coefficient

## 1 Introduction

Diabetes is a universal chronic disease around some developed countries and developing countries. The diabetic eye disorder is called as diabetic retinopathy; it is a medical state in which high amount of sugar level causes harm to retina. It is foremost origin of blindness. The chances of occurring diabetic retinopathy is upto 80% if the person suffering from diabetes mellitus from 20 years or more. DR causes major damage to blood vessels of the retina due to contain of high amount of sugar in it. The main prodrome can happen to patient is blur or deficiency in the vision, and it is a major cause of blindness. The key symptom of DR is exudates. Normally to detect the exudates caused to retina, we need to go to ophthalmologists test the

---

A. Upadhyay (✉) · P. Kantelia · R. Parmar

Thakur College of Science and Commerce, Kandivali (E), Mumbai 400101, India

e-mail: [anandhari6@gmail.com](mailto:anandhari6@gmail.com)

P. Kantelia

e-mail: [kanteliap@gmail.com](mailto:kanteliap@gmail.com)

R. Parmar

e-mail: [rohan260499@gmail.com](mailto:rohan260499@gmail.com)



**Fig. 1** Vision of normal eye and DR affected eye [3]

pupil dilation with chemical solution which consume lots of time. So by making use probabilistic neural network algorithm, we can recognize defect in eye. The PNN algorithm has been tested on an image database and compared with the performance of a human eye. The pervasiveness of diabetic retinopathy among individuals those who are suffering from diabetes is around 28.5% in United States, and in India, it is 18% which is a major problem among the country [1]. The primary problem of an individual diagnosed with diabetes that is high amount of glucose level which affect various organs like brain, nerves, kidney and evidently eyes.

Diabetic retinopathy recognition is becoming challenging as the time human reviewer presents their evaluation, on periodic a day or later; thus, the retard assessment leads to lost of follow-up, lack of treatment and misinterpretation. Vascular abnormalities and by the existence of lesions in it, clinicians can spot diabetic retinopathy very well. This process is effectual, but the resources are high in demand. There is shortfall of prowess and tools essential in this areas where the demand of native inhabitants who are suffering from diabetes is high and detection of DR is mostly required. The requirement of a multidisciplinary and computerized technique for the recognition of DR screening, and prior efforts have made quality advancement by implementing pattern recognition, image classification and machine learning [2] (Fig. 1).

## 2 Literature Review

Darshit Doshi, Aniket shenoy, Deep Sidhpura and Dr. Prachi Gharpure have performed and implemented deep convolutional neural network (DCNN) for the recognition of diabetic retinopathy. With the help of the deep learning algorithm, they have tested and implemented the detection of various stages of disease in computerized way in their paper. To process the high-resolution images and classify them automatically into five different stages of disease on the bases of their severity, they first designed and then implemented the DCNN with the help of GPU acceleration.

They got the accuracy of 0.386 on kappa value on the single model with convolutional neural network and 0.3996 on three similar models [4].

Mohit Singh Solanki has made the project in an attempt towards finding an computerized way of detecting diabetic retinopathy disease in its premature stage. They have implemented supervised learning methods to classify the image dataset into five classes. They have used 500 images to train and same number of images to test, and they got accuracy of 55%, which took 6–7 h for the neural network to process those images. For this task, they used techniques of image processing and filters to enhance many important features and then using neural network for classification [3].

### 3 Algorithm

**Probabilistic Neural Network (PNN):** PNN is used in classify the problem recognition of the image when the input is present. Probabilistic neural network is used to detect the image of an infected eye.

**Input Layer:** Every neuron is represented as predictor variable in the input layer. There are N number of classes which are used by neurons. Hidden layers neurons get the values from the input neuron.

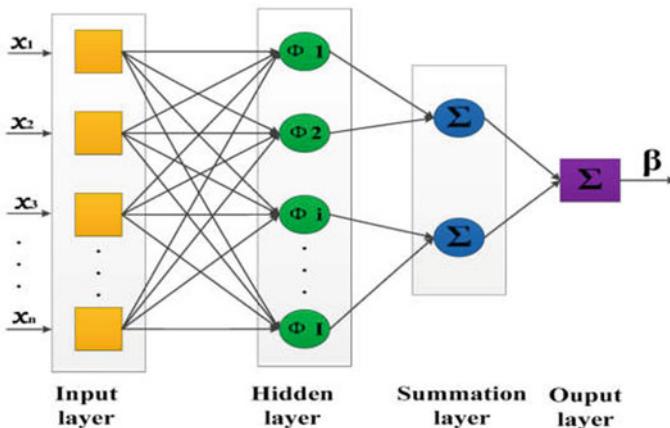
**Pattern Layer:** For each case of training dataset, pattern layer contains single neuron.

**Summation Layer:** In PNN, for the network, single pattern neuron is presented in category to the target variable. Each case is trained, and target category is kept in each hidden neuron. The gain of a hidden neuron is the pattern to correspond to the hidden neuron category, and the image neuron adds the values for the class they represents.

**Output Layer:** This layer differentiates the votes for target class, which is there in pattern layer and most votes are used to predict the targeted class (Fig. 2).

### 4 Dataset

For this work, the data is collected from Kaggle [5]. Kaggle image database contains more than 80,000 images with high resolution. As PNN occupies lots of memory to process such high-resolution images, so we took around 100 images of healthy and diabetic eye and we resized the images to  $250 \times 250$  resolution to reduce the computational complexity. For training purpose, we have cropped the images which contain normal and diabetic patches. We used more than 150 images to train our neural network and prepared subset of dataset of two classes as show in Table 1. Multiple models and various types of cameras under different radiance are used to derive images for the dataset.



**Fig. 2** Architecture of PNN

**Table 1** Error matrix

Class	Affected	Non-affected	Total	Procedure accuracy (%)
Affected	2222	104	2326	95.53
Non-affected	13	1115	1128	98.84
Total	2235	1219	3454	
User accuracy (%)	99.41	91.46		96.61

$$\text{Accuracy} = (3337/3454) * 100$$

$$\text{Accuracy} = 96.61\%$$

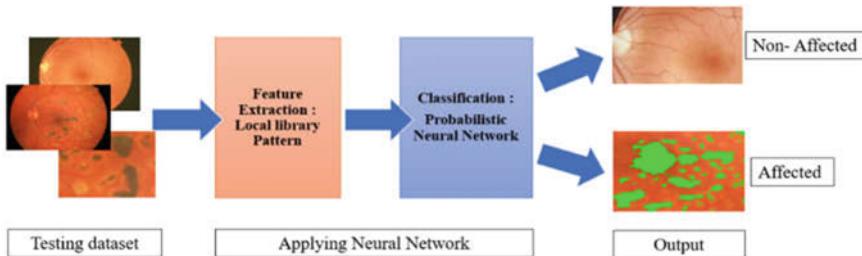
## 5 Methodology

In the first stage, the image characterizations are executed using local patterns considering rotation invariant, uniform patterns. Then, a set of features is provided to new PNN which is inbuilt function in MATLAB for PNN to perform the classification task (Fig. 3).

To classify the fundus image as healthy or diabetic, we propose a method that divides the process into two stages :

### 5.1 Extraction of Features

Extracting pertinent information (features) in an image is the common objective of image classification, so in this stage, we provided image as an input to the neural network and extraction of feature takes place. Firstly, we converted the target class indices to vectors with spread value of 25. In this process, the image vectors are



**Fig. 3** Stages of the proposed technique

reshaped into RGB matrix, respectively, and then it creates the network and simulates it. Vec2ind function is used to convert output vectors into row or indices, i.e., instead of process the whole image as a high-dimensional vector, the idea is to describe only local features of an object.

## 5.2 Classification of Image

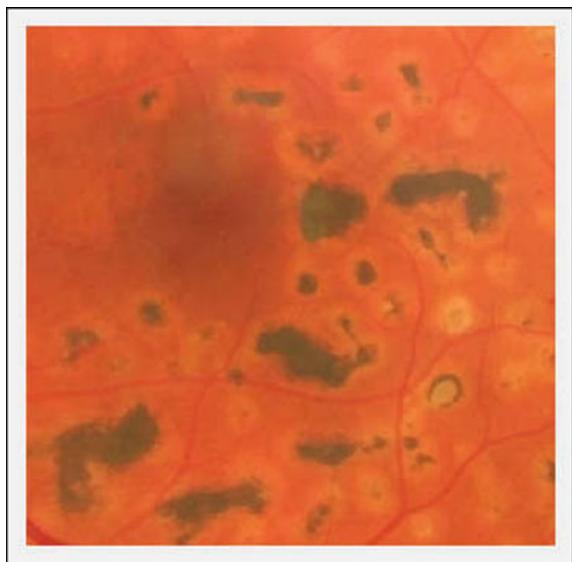
In further, vectors of image are reshaped and then the neural network process the input image pixels with dataset that contains the RGB values of both the classes. The affected area is classified as class 1 and non-affected area as class 2. To perform this local feature extraction, methods called confusion matrix and kappa coefficient are used.

## 6 Result

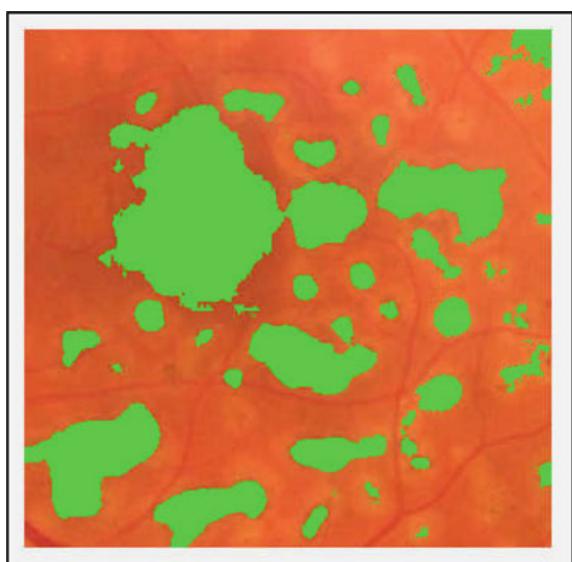
Proposed PNN build method is tested with varied test image specimen. This approach is used to evaluate the performance in terms of percentage accuracy, color image processing (RGB). We have collected 3454 (RGB) values which is divided in 2027 values in class 1 and 1427 values in class 2. After applying the neural network, we got following results.

We have designed, trained and implemented probabilistic neural network algorithm to detect the diabetic retinopathy, and we trained our dataset with small subset of images and labeled them with two classes. The defected areas are highlighted with the green color patches. The dark spots are the infected areas which can be seen in diabetic patients which is shown in Fig. 4, and when we applied the PNN, the infected areas got highlighted with green color which is shown in Fig. 5. This shows the result and the area which is affected by diabetic or not.

**Fig. 4** Diabetic retinopathy image before applying algorithm



**Fig. 5** Diabetic retinopathy image after applying algorithm



## 7 Result of Confusion Matrix and Kappa Coefficient

**Confusion matrix:** Confusion matrix is also known as error matrix. Error matrix is a summary of prediction to classify the outcomes on the following based problems. Various accurate and inaccurate projections are done by enumerating values and

**Table 2** Accuracy evaluation

S. No.	Criterion	Gain (%)
1	Accuracy	96.61
2	Kappa gain	0.9246

disintegration of individual class, and it leads to confusion matrix. Confusion matrix gives the intuition about the type of error being occurred and falacy made by classifier.

**Kappa coefficient:** Cohen's kappa coefficient is a statistic that is used to measure inter-rater reliability for qualitative items. Equation for calculating kappa value is stated below. If the value of Kappa ( $K$ ) is 1 or nearer to 1 than it is more accurate, whereas if the value of  $K$  is near 0 than it is less accurate and they are labeled.

$$\hat{K} = \frac{N * \sum_{i=1}^k x_{i=1} - \sum_{i=1}^k (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^k (x_{i+} * x_{+i})}$$

**K = 0.9246 (very good)**

We have compared the Kappa coefficient score and accuracy of the confusion matrix from which we got very good score in kappa as results are shown in the following table. In Table 1, confusion matrix represents the accuracy of user's and procedure and also shows amount of data of both the classes, i.e., affected and non-affected. Table 2 shows actual results.

## 8 Conclusion

In this work, we have used probabilistic neural network which is feed forward neural network to detect retinopathy in retina image and to get the area affected with diabetic disease. We have implemented this algorithm in MATLAB to detect affected area and mark with appropriate color. After many trials and errors, we finally designed the neural net. The trained neural network accuracy has been found is 96.6% in recognizing the diabetic area. Due to less configuration of machine, we were unable to process large amount of dataset. But still, we were able to achieve maximum accuracy with small size of images. As PNN is slower than other perceptron network which requires much more memory space to process the neural network on the system.

## 9 Future Direction

Our future goal or work is to increase size of test and training data set to recognize and make more accurate result with the help of system with higher configuration, so that we can provide high-resolution images to get highly precision solutions. Also

it's going to help patients who are suffering from diabetic retinopathy to easily get the report of area, which is infected by DR after providing the fundus image as input to the algorithm and to get the results on the go.

**Acknowledgements** We thank Kaggle [5] for free dataset online.

## References

1. Sopharak A, Uyyanonvara B, Barman S, Williamson TH (2008) Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods. *Comput Med Imaging Graph* 32(8):720–727
2. Schaefer G, Leung E (2007) Neural networks for exudate detection in retinal images. In: International symposium on visual computing. Springer, Berlin, Heidelberg, pp 298–306
3. Solanki MS (2015) Diabetic retinopathy detection using eye images. Artif Intell Course Project 1–10
4. Doshi D, Shenoy A, Sidhpura D, Gharpure P (2016) Diabetic retinopathy detection using deep convolutional neural networks. In: 2016 International conference on computing, analytics and security trends (CAST). IEEE, pp 261–266
5. <https://www.kaggle.com/c/diabetic-retinopathy-detection>
6. De la Calleja J, Tecuapetla L, Medina MA, Bárcenas E, Nájera ABU (2014) LBP and machine learning for diabetic retinopathy detection. In: International conference on intelligent data engineering and automated learning. Springer, Cham, pp 110–117

# Application of Unscented Kalman Filter for Parameter Estimation of Nonlinear Systems



Urmila Solanki, Ganesh P. Prajapat, and Manoj Chhimpa

**Abstract** Sometimes the parameters of a system dynamics are not exactly known while these are required to set the control law and update the existing control scheme. This becomes much difficult when the dynamics of the system is nonlinear. Thus, this paper deals with the estimation of the parameters of a nonlinear system using unscented Kalman filter (UKF). The UKF handles the nonlinear dynamics without linearization and approximation during the estimation and hence estimates the parameters as well as states perfectly. A well-known example of nonlinear dynamics, Van Der Pol oscillator system, has been used to illustrate the parameter estimation. The simulation of the Van Der Pol oscillator has been done first to generate the measurements, and then, the state and measurement model of the system have been setup which further have been used during the estimation.

**Keywords** Nonlinear system dynamics · Van der pol oscillator · Unscented Kalman filter (UKF) · Parameter and state estimation

## 1 Introduction

The mathematics description of the dynamical nonlinear system involves unknown variables of higher degree polynomial or unknown functions when system equations appear as differential equations. The response of the dynamics of these systems and

---

U. Solanki · G. P. Prajapat (✉) · M. Chhimpa

Department of Electrical Engineering, Government Engineering College, Bikaner,  
Rajasthan 334004, India

e-mail: [prajapat2008@gmail.com](mailto:prajapat2008@gmail.com)

U. Solanki

e-mail: [urmilarsolanki@gmail.com](mailto:urmilarsolanki@gmail.com)

M. Chhimpa

e-mail: [manoj.5782kumar@gmail.com](mailto:manoj.5782kumar@gmail.com)

the values of the unknowns are difficult to evaluate. The approximate linearization not gives the accurate solution for the nonlinear system. The true values of nonlinear system states and parameters are needed for the designing of the control scheme for these systems.

In the phenomena of states, space of the nonlinear system with random Gaussian noise presented by the stochastic nonlinear model uses the Kalman filter [1]. At present, the extended Kalman filter (EKF) is widely used for the state and parameter estimator for nonlinear system, although it has some flaws as it uses standard linearization methodology and susceptibility to bias and deviation in the estimation and requires sufficient differentiability of dynamics of system states. Another derivative-free replacement to EKF is unscented Kalman filter (UKF) [2–4]. There are some points which show the benefits of using unscented Kalman filter over the other estimation approaches for the state and parameter estimation of the nonlinear system, explained below:

- The advantage of unscented Kalman filter algorithm is that no linearization step present in the estimation process through [3, 5–7].
- The UKF is employed to extract the internal state parameters of the power system model as all the state parameters are not available.
- The methodology of UKF can be efficiently used in real time applications.
- The UKF used for solving problems related to estimation is an effective discrete-time recursive filter.
- It has properties over EKF that it is easy to implement, more accurate, and use same order of calculation of linearization.
- The UKF approach definitely exceeds in terms of accuracy and smoothness comparing the state estimation in the presence of noise.
- The UKF estimates the states of nonlinear dynamical systems using noise-free as well as noisy electrical measurements with its properties of robustness and speed of convergence [8].

In this paper, UKF is used for the estimation of the scalar parameter of the Van der pol oscillator which indicates the nonlinearity and strength of the damping.

## 2 Studied Nonlinear System Dynamics

The common example considers for the study of the dynamics of nonlinear system is Van der Pol oscillator which is used for the designing of various mechanical, electrical, and laser oscillatory systems. The wide range of applications of the Van der Pol oscillator are as chemical oscillations, electric circuit, heartbeat, circadian rhythm, and biological rhythm which show its nonlinear oscillatory behavior [9–11].

The following second-order nonlinear differential equation shows the mathematical model Van der Pol oscillator

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0 \quad (1)$$

where

$x$  presents the coordinate position and it is a function of time.

$\mu$  a non-vector parameter which is responsible for the nonlinearity of the system and strength of damping.

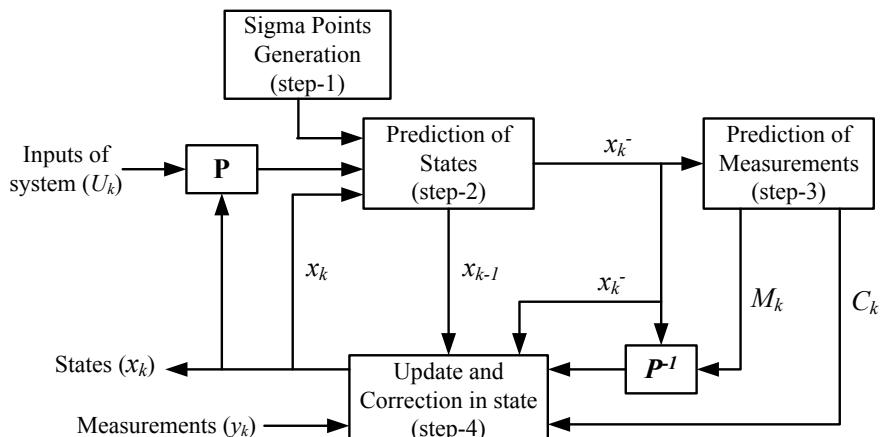
When  $\mu = 0$ , no damping element present in the system then it becomes simple harmonic oscillator, and there is an action of the energy conservation always take place.

When  $\mu > 0$ , system has unforced oscillation. The solution of the  $\dot{x} = 0$  has a limit circle and the position of this circle shows the nature of the system as near to origin unstable system and far from origin damped system.

### 3 Unscented Kalman Filter

The unscented Kalman filter can give the accurate second-order Taylor expansion and superior performance with the same computational complexity without linearization and Jacobian solution. UKF is an iterative algorithm; therefore, it can be used in real practices efficiently. The unscented Kalman filter is an uncomplicated supplement of the unscented transformation (UT) to the recursive estimation. The UT is based on the instinct that it is simple to be close to a probability distribution than it is for approximation and transformation of a random nonlinear function.

The algorithm for the estimation of states and parameters through UKF involves four steps [1] which can be understood with the block diagram as shown in Fig. 1 and its explanation.



**Fig. 1** Block diagram showing the steps of UKF

**Step-1:** The sigma points, which are the set of  $(2N + 1)$  vectors, can be generated by using the mean value of the states,  $\bar{x}$  and matrix of covariance,  $P$  with the following equations:

$$x_0 = \bar{x} \quad (2)$$

$$\begin{aligned} x_i &= \bar{x} + \left( \sqrt{(N + \zeta) P_{k-1}} \right)_i, \\ i &= 1, \dots, N \end{aligned} \quad (3)$$

$$\begin{aligned} x_i &= \bar{x} - \left( \sqrt{(N + \zeta) P_{k-1}} \right)_{i-N}, \\ i &= N + 1, \dots, 2N \end{aligned} \quad (4)$$

$$\zeta = \alpha^2(N + \nu) - N \quad (5)$$

The weighted sample calculated from the following Eqs. (6)–(8) are used in the steps of this algorithm where the prediction and correction of the state and measurement are taking place which are further used to determine the state vector and matrix of covariance.

$$W_0^m = \frac{\zeta}{N + \zeta} \quad (6)$$

$$W_0^c = \frac{\zeta}{N + \zeta} + (1 - \alpha^2 + \beta) \quad (7)$$

$$W_i^m = W_i^c = \frac{1}{2(N + \zeta)} i = 1, \dots, 2N \quad (8)$$

where  $N$  is the total number of states, and here, it is equal to 2 for the Van der pol oscillator considered as example of nonlinear system,  $\zeta$  is a scaling factor,  $\alpha$  is the deciding factor for the sigma points spread around the mean value of the states ( $\bar{x}$ ), and usually, it is set to 0.001,  $\nu$  is secondary scaling factor (usually set to 0), and  $\beta$  is the previous knowledge of the state distribution around the mean value of the states, and its typical value for Gaussian distributions is 2.

**Step-2:** The sigma points generated from step-1 has been distributed along with the discrete nonlinear function,  $f$  to generate the sigma points of predicted states,  $x_k^i$ .

$$x_k^i = f(x_{k-1}^i, z_{k-1}, U_k) \quad (9)$$

The predicted mean state vector,  $x_k^-$ , and the predicted matrix of covariance  $P_k^-$  can be obtained by using the sigma points of predicted states and weighted samples as follow:

$$\bar{x}_k = \sum_{i=0}^{2N} W_i^m x_k^i \quad (10)$$

$$\bar{P}_k = \sum_{i=0}^{2N} W_i^c \left( (x_k^i - \bar{x}_k) (x_k^i - \bar{x}_k)^T \right) \quad (11)$$

**Step-3:** The predicted mean state vector obtained from Eq. (10) is used to generate the sigma points of predicted measurement in Eq. (12) by using the nonlinear measurement function,  $\mathbf{h}$ , and the predicted measurement vector,  $\bar{y}_k$ , is obtained by using the sigma points of predicted measurement and weighted samples.

$$y_k^i = \mathbf{h}(x_k^i, z_k) \quad (12)$$

$$\bar{y}_k = \sum_{i=0}^{2N} W_i^m y_k^i \quad (13)$$

Further, the covariance matrix of measurement,  $M_k$ , and the cross-covariance matrix of state and measurement,  $C_k$ , are computed using  $R$ , as

$$M_k = \sum_{i=0}^{2N} W_i^c \left( (y_k^i - \bar{y}_k) (y_k^i - \bar{y}_k)^T \right) \quad (14)$$

$$C_k = \sum_{i=0}^{2N} W_i^c \left( (x_k^i - \bar{x}_k) (y_k^i - \bar{y}_k)^T \right) \quad (15)$$

These two matrixes in (14)–(15) are used for Kalman gain update and state correction.

**Step-4:** Finally, the Kalman gain,  $K_k$ , the updated estimated states,  $x_k$ , and the matrix of covariance,  $P_k$ , are determined with the help of the results obtained from step-2 and step-3, where  $y_k$  are the obtainable measurements at  $k$ th time instant.

$$K_k = C_k M_k^{-1} \quad (16)$$

$$x_k = \bar{x}_k + K_k (y_k - \bar{y}_k) \quad (17)$$

$$P_k = P_k^- - K_k M_k K_k^T \quad (18)$$

At the end of the estimation process using UKF, the estimated state was obtained from Eq. (17) where the damping parameter,  $\mu$ , has also been considered as state with its derivative equal to zero.

## 4 Proposed Scheme of Parameter Estimation

The proposed work starts with the setting up of the mathematical model of the Van Der Pol oscillator system. The system state model has been made as follows:

$$\dot{x}_1 = x_2 \quad (19)$$

$$\dot{x}_2 = \mu(1 - x_1^2)x_2 + x_1 = 0 \quad (20)$$

while the measurement model has been made out of the two states of the system. However, any realistic model of the measurement can be used but it should be such that it depends on the states as well as parameter  $\mu$  of the system dynamics.

The UKF algorithm uses the state model of the system dynamics only, and thus, one has to define the constant parameter as a state. Hence, the state equation of the damping parameter,  $\mu$ , has also been considered as state with its derivative equal to zero as follows

$$\dot{\mu} = 0 \quad (21)$$

Ultimately, the system made up with the three states (two from the oscillator and one of parameter) has been directly used in the UKF algorithm which are as follows

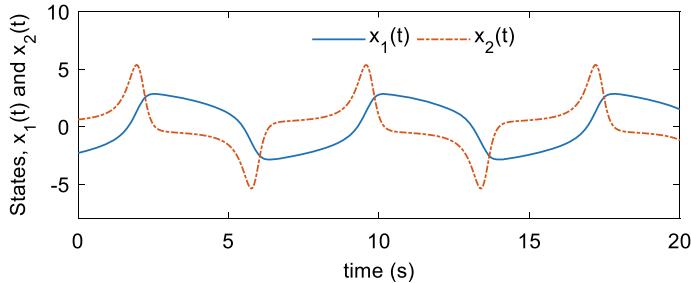
$$x = [x_1 \ x_2 \ \mu]^T \quad (22)$$

## 5 Simulation Results

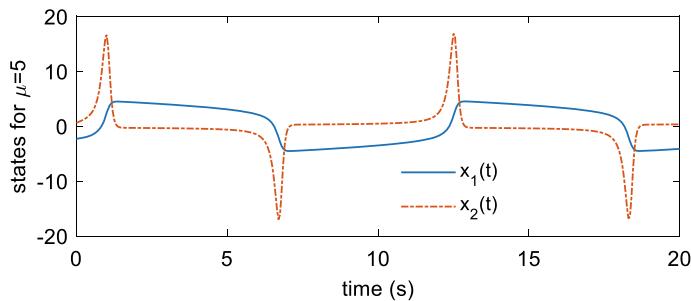
The simulation of the system has been performed in MATLAB using nonlinear solver, *ode15s*, considering the nonlinear differential equation model of the system dynamics detailed in (19)–(20). The initial guess of states during the solution by ode solver has been found using *fsolve* function in MATLAB.

### 5.1 *Simulation for Measurements*

The nonlinear system modeled in the section above has been simulated using *ode15s* considering two states as their derivatives are available as in (19)–(20). The simulation has been run for the time span of 20 s with the initial guess found by *fsolve* which exactly gives the equilibrium point at that instant. The *ode15s* solver is a stiff solver, and it is capable to handle the nonlinearity of continuous function efficiently. It is also capable to solve an algebraic differential equation (DAE) problem. The solver



**Fig. 2** States  $x_1(t)$  and  $x_2(t)$  for  $\mu = 2$



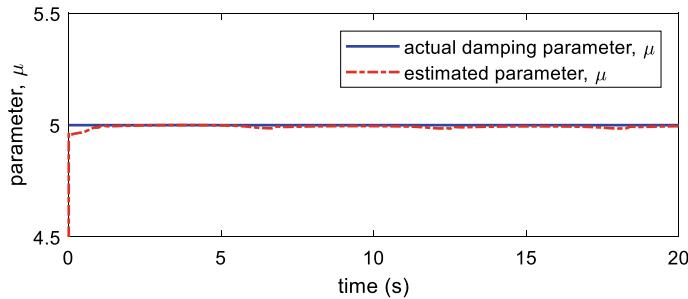
**Fig. 3** States  $x_1(t)$  and  $x_2(t)$  for  $\mu = 5$

returns to a column vector of two states with the time. Solution found in such a way is drawn in and as shown in Figs. 2 and 3 for the different values of the damping parameter,  $\mu$ . The nonlinearity of the oscillator can be clearly seen in the figures. It can also be seen from the Figs. 2 and 3 that the frequency of the Van Der Pol oscillator depends on its damping coefficient,  $\mu$ . The higher the value of  $\mu$ , the frequency is lower and vice versa. This is verifying the perfection of the simulation using ODE solver.

Further, the measurements have been extracted from the simulation using these states. For the purpose of simplification and better illustration of the parameter estimation, the states themselves have been used as measurements. The measurements found in this way have been saved with a time step of 0.02 s and further have been used in the UKF algorithm.

## 5.2 Estimation of Parameters

Once the state model has been setup along with the measurement model with measurements, the UKF algorithm directly uses them for the estimation. The algorithm has been run after its initialization with the covariance matrix and states. In



**Fig. 4** Estimated damping parameter,  $\mu$

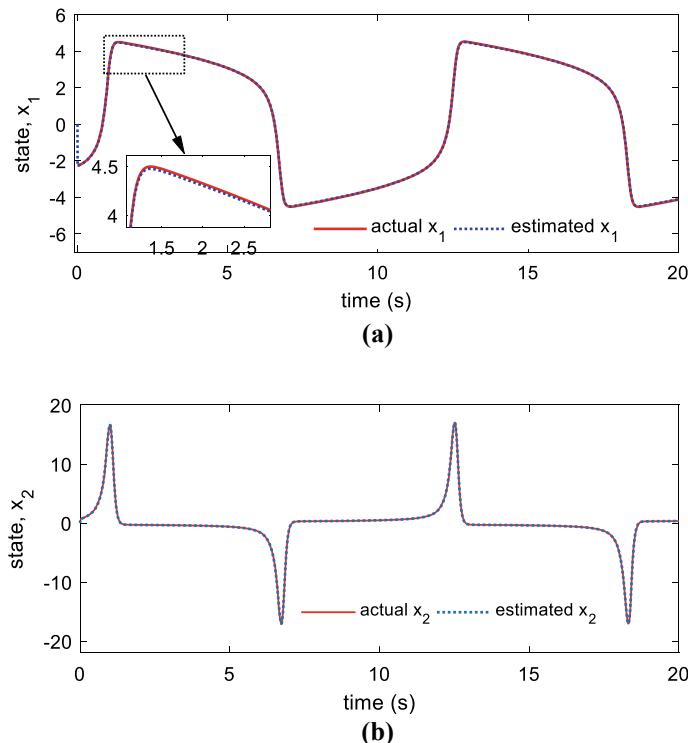
every step of run, the UKF takes the initial value of states, measurements, create sigma points, pass them through state model, and then measurement model to form the Kalman gain which finally gives estimate the states. The damping parameter of the Van Der Pol oscillator is also treated as state and it comes out of algorithm as state and gets updated every iteration. This is something like the online estimation of the parameter, and thus if there is a change in the parameter in between, the updated parameter will come out of the UKF.

The estimated damping parameter,  $\mu$ , is shown in Fig. 4 where it can be observed that the UKF estimates the parameter perfectly. Estimation can be made more precise by the fine-tuning of the process and measurement covariance matrixes.

Further, the states have also well estimated as their response is as same as actual one as shown above in Fig. 5a, b. Furthermore, the estimation has also been performed at the different values of the damping parameter,  $\mu$ , and observed that higher its value, lower the estimation error. Same can be seen from the Table 1 where the estimation error is 2.35% for  $\mu = 2$  while it is only 0.25% for  $\mu = 10$ .

## 6 Conclusion

The information of a parameter of the nonlinear system dynamics is very difficult to know, and hence, it needs the estimation. Thus, this paper illustrates the estimation of the parameters of a nonlinear system using unscented Kalman filter (UKF). The UKF handles the nonlinear dynamics with a significant accuracy even with the noisy measurements and estimates the parameters as well as states perfectly. A well-known example of nonlinear dynamics, Van Der Pol oscillator system, has been used to illustrate the parameter estimation. The simulation of the Van Der Pol oscillator has been done first to generate the measurements, and then, estimation has been performed. It was observed that the UKF estimates the parameter,  $\mu$ , perfectly along with the states. It is also concluded from the results that the estimation error is much lower when the value of parameter is higher.



**Fig. 5** Estimated States, **a**  $x_1(t)$  and **b**  $x_2(t)$  for  $\mu = 5$

**Table 1** Estimation Error with Different Values of Parameter,  $\mu$

Actual value of parameter, $\mu$	Estimated $\mu$	% Estimation error
2	2.047	2.35
5	5.036	0.72
10	10.025	0.25

**Acknowledgements** This research work is funded from the Collaborative Research Scheme (CRS) of RTU (ATU) under TEQIP-III.

## References

1. Julier SJ, Uhlmann JK (2004) Unscented filtering and nonlinear estimation. Proc IEEE 92(3):401–422
2. Yazdanian M, Mehrizi-Sani A, Mojiri M (2015) Estimation of electromechanical oscillation parameters using an extended kalman filter. IEEE Trans Power Syst 30(6):2994–3002

3. Ghahremani E, Kamwa I (2011) Online state estimation of a synchronous generator using unscented Kalman filter from phasor measurements units. *Energy Convers IEEE Trans* 26(4):1099–1108
4. Azad SP, Tate JE (2011) Parameter estimation of doubly fed induction generator driven by wind turbine. Phoenix, AZ, pp 1–8
5. Wan EA, Van Der Merwe R (2002) The unscented Kalman filter for nonlinear estimation. In: Proceedings of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium (Cat. No.00EX373), pp 153–158 (2002)
6. Simon D (2006) Optimal state estimation: Kalman,  $H\infty$ , and nonlinear approaches. Wiley, New Jersey
7. Yu S, Emami K, Fernando T, Ju HHC, Wong KP (2016) State estimation of doubly fed induction generator wind turbine in complex power systems. *IEEE Trans Power Syst* 31(6):4935–4944
8. Lalami A, Wamkeue R, Kamwa I, Saad M, Beaudoin JJ (2012) Unscented Kalman filter for non-linear estimation of induction machine parameters. *IET Electr Power Appl* 6(9):611–620
9. Kanamaru T (2007) Van der Pol oscillator. Scholarpedia
10. Tsatsos M (2008) The Van der Pol Equation. arXiv0803.1658: ArXiv Prepr arXiv0803.1658
11. Ginoux JM, Letellier C (2012) Van der Pol and the history of relaxation oscillations: Toward the emergence of a concept. *Chaos*

# Question Answering System Using LSTM and Keyword Generation



Minakshi Tomer and Manoj Kumar

**Abstract** Question answering system is an area under natural language processing and information retrieval which automatically answers the questions generated by humans. This work represents an approach for building a system that generates answers for the question based on deep learning neural network which has the competence of processing the information present inside the dataset and enables the user to obtain an insight from the SQuAD dataset by inviting questions in natural language form. Key stages of this approach cover corpus pre-processing, question pre-processing, answer generation, deep neural network for answer extraction and keyword generation. The concept of keyword generation is a novel idea implemented to enable naïve user of the system to apprehend the passage. The system is competent in interpreting the question, responding to the user's query in natural language form along with generating the keywords. The performance was measured on SQuAD dataset using EM and F1 score.

**Keywords** Question and answer · Artificial neural network · Sequence-to-sequence model · Recurrent neural network · Long short-term memory (LSTM) · Word embedding

## 1 Introduction

The job of question answering (QA) and machine comprehension (MC) within the natural language processing and computer vision societies have achieved meaningful popularity across years. Over various tasks in the text and image fields, systems

---

M. Tomer (✉)

University School of Information Communication and Technology, GGSIPU, Delhi, India  
e-mail: [tomer.minakshi@gmail.com](mailto:tomer.minakshi@gmail.com)

Maharaja Surajmal Institute of Technology, Delhi, India

M. Kumar

Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India  
e-mail: [manojkumar@aiactr.ac.in](mailto:manojkumar@aiactr.ac.in)

trained end-to-end presently produce likely results. The use of neural attention mechanism has been one of the principal determinants to the progress that find a suitable content inside a paragraph (for MC) which allows the system to concentrate on a particular region, that is most suitable to respond to the query [1, 2].

Attention mechanisms in preceding works had generally one or more of the resulting features. Firstly, summarize the text into a vector of fixed length using word embedding to answer the question by extracting the most relevant information, and the expected attention weights are often utilized. Secondly, they are usually temporally progressive in the text domain, whereby at the preceding time level the attention weights at the prevailing time step are a function of the accompanied vector. Third, they are frequently unidirectional, wherein each query characterizes on the meaning paragraph or the memory.

QA system research aims to dispense with a broad diversity of question kinds including list, theoretical, definition, what, why, fact, how, and cross-lingual questions. Pretext of the question answering system paradox, the question answering system can be classified into closed domain system and open-domain systems. Closed domain question answering systems fall under a definite area and can be an indulgent task because NLP operations can exploit area-specific information often formalized in ontologies. Alternatively, closed domain may refer to a condition where restricted kind of questions is acquired, such as questions accounting for representative preferably than procedural knowledge. Open-domain question answering can simply rely on world knowledge and usual ontologies. Such systems normally have much more access to the data from which the answer can be generated.

## 2 Literature Survey

Vast research in the area of Q&A domain provides a clear structure for this model. This prompted us taking a deep learning-based approach. Artificial intelligence with its subfields machine learning and deep learning has been earning plenty of eminence in the present era. Natural language processing, speech recognition, visual object recognition, convolutional neural networks [3] etc.

Question answering: the answer to some fact-based question is a line span in the article [4] many novel discoveries have occurred in the area of QA from KBs with the formulation of resources like Simple Questions [1] the Freebase KB [5] and Web Questions [6] or on automatically selected KBs, e.g., Open IE triples and NELL [7]. In context of study in Natural Language Processing field, most of the research are being preferred in Question Answering domain. QA is most generally employed in systems that simulate human communication such as chatbots and other implementations. Tried and tested ways of parsing from the domain of NLP research are used.

Task transfer and multitask learning [8] have a deep history in machine learning, as well as in NLP [9]. Several works have tried to connect multiple QA training datasets via multitask learning to (i) provide a single general system accomplished of asking

various sorts of questions due to the unavoidably complex data distributions across the source datasets and (ii) gain improvement across the datasets via task transfer. The goal of this paper is to examine how so distinct systems can work in an open-domain QA structure.

Information retrieval [10], TF-IDF, and n-gram overlap of sentence [11], multi-modal mobile clues [12], extraction techniques consisting of distant supervised learning [13] are some of the techniques that were surveyed.

Augmented neural networks [14–16] and launch of further training and evaluation datasets like SQuAD [17], Wiki Reading [18], or Quiz Bowl [19] and CNN/Daily Mail based on news reports [20].

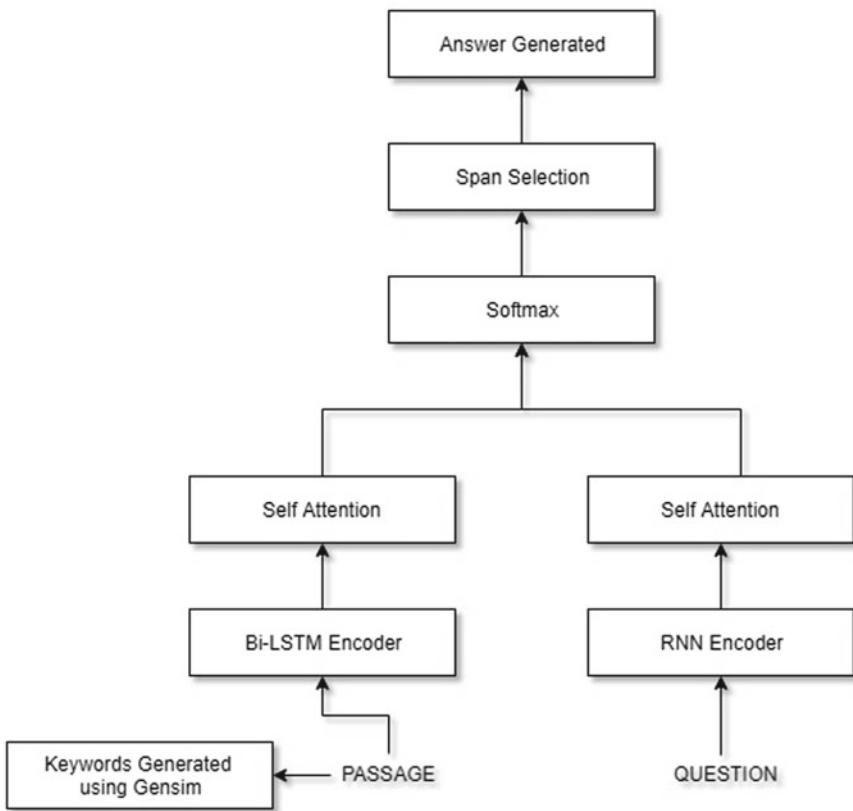
To encode text and query, respectively, it uses Bidirectional RNN [21, 22] and uses query description to match with every token from the paper. For factoid RCQA [20], Attentive Reader was the primary neural paradigm. Attention Sum Reader [23] examines the model to just predict positions of the accurate answer in the record and the training rate and test accuracy are both hugely improved on the CNN/Daily Mail dataset. Cui et al. [24] also analyzed Attentive Reader and delivered a higher performance.

RNN encoders are not utilized by Window-based Memory Networks [25] which is displayed along with the CBT dataset [15] but sets contexts as memory and embedded contexts are matched with questions. The mechanism of the model is to determine the similarity between answer context with question representation. In contrast, memory improved neural networks like Neural Turing Machines [16] and its alternatives [26] was also a potential candidate for the job, and [26] detailed results on the bAbI task, which is more acute than memory networks. Likewise, sequence-to-sequence models were also used [27], but the outcomes were not satisfying.

Benchmark datasets perform a crucial part in the recent growth in question answering analysis and reading comprehension. Current datasets can be divided into two types: according to whether they are manually specified. Those that are marked by individuals are invariably in great quality [28–30] but are way too small for training advanced data-intensive paradigms. The commonly occurring data that can be automatically generated can be very large [15, 20] which enable the training of more expressive paradigms, although the aim is to predict the missing information an entity in a paragraph. As determined by Cui et al. [24] the CNN/Daily News dataset [20] requires shorter reasoning than thought, and reason that efficiency and performance is substantially saturated. From our literature survey, we have established that the QA domain is an emerging area among researchers and several distinct methods have been carried out to deal with this area and we have concluded that using a question–answer system using the SQuAD dataset gave a comparable results. The scope has been thus constricted to improve the precision of answers presented.

### 3 Methodology

This model employs SQuAD [17] dataset for the training and testing of deep neural network model which has the capability to answer user queries from the provided passage. Proposed model of the system architecture is presented in Fig. 1. The pre-processed data are fed into the neural network to train the system and it generates a model based on machine learning. Additionally, the inputs which are a set of passage along with questions are also being processed using word embedding techniques and answers which are generated as an output is based on deep neural network techniques (LSTM). The output generated using LSTM is more human friendly.



**Fig. 1** System architecture

### 3.1 *Corpus Pre-processing*

To create the model, one of the primary tasks that need to be performed initially is data pre-processing which includes stop word removal, stemming, and tokenization. Originally, the corpus available is an unorganized text content, which is easily interpreted by humans but not by machines. Semantic parsing can be done by using word embeddings, to obtain the sense from text to enable NLU. To understand the contextual relationship of words, for a language model to learn the meaning of a word, it needs.

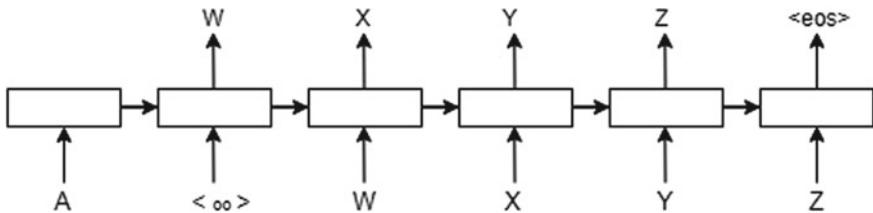
Batching is the process of separating the training dataset into multiple batches. These batches are used as input to train the network where the weights are modified once each batch has been utilized by the network. As an example, training set with 87,000 question-answer pairs is further divided into batches and these batches are then used during training. For training the neural network model, a dataset of 32 batches is used at a time in this paper. The performance and efficiency of the training method is enhanced as compared to the other online training process. We choose to use this process because of the lack of hardware devices and time restrictions. To encode, we use a Bidirectional LSTM.

During the training process utilizing complete dataset, an epoch is attained and the process is redone with the same dataset. The model will overfit if this is performed too much. Initially, the proposed model was trained over 30 epochs in this model. However, while increasing the epochs results were getting better. Therefore, the epochs were set to be 40 for this model.

### 3.2 *Question Pre-processing*

The question asked by the user has to be pre-processed as well simultaneously with the additional text data. This is done using a simple recurrent neural network, where word embeddings is applied to generate a vector for each word. The token from the questions is mapped to its corresponding vector space using embedding layer, which preserves the context similarity among words. A common pre-trained word embedding model known as GloVe is used as the embedding layer in the question processing segment.

Sequence-to-sequence model with LSTM-based encoder/decoder cells is used for training the question sequences. The LSTM encoder is used to encode the question sequence into a vector of fixed length using Pytorch framework.



**Fig. 2** An RNN cell

### 3.3 Deep Neural Network for Answer Extraction

The main component primarily functions as the portion where the actual decision making happens. The structured data present in the pre-processed corpus is to produce an answer to the pre-processed question which is the primary aim of this module. In this research, to obtain the answers from the dataset, we use a recurrent neural network (RNN). In previous researches, on the comparison, it was found that, when it comes to text processing, RNN's function much better than other available neural networks. The basic RNN architecture is presented in Fig. 2.

LSTMs are being used because they have some memory capacity and are better at working with vanishing gradients. We are utilizing three LSTM layers with 128 hidden units. This was thought as the best approach after understanding research papers.

The model receives its inputs from the question pre-processing component and generates an output response into the answer generation component so that a human-readable result can be generated.

### 3.4 Answer Generation

Since the dataset does not contain the answer in the required format, hence this component is used. The difficulty with producing natural language from a particular source of text is that it is the backward method to what is taken by maximum other researchers. Deep neural networks are used as predictive paradigms and are used to recognize patterns.

Once the question is encoded and passed as an input into the LSTM cells, the neural network will process the output and generates the answer that is placed in the entered passage by the user. This provides us with a collection of words that the neural network conceives as the answer to the question.

### 3.5 *Keywords Generation*

In order for the user to get a brief idea about the passage, we incorporated Genism's keyword functionality which returns so of the keywords from the passage. Generating important words from the passage was aimed to help the user to ask questions by having context from the passage.

## 4 Results and Discussion

The system enables the user to raise similar questions over the dynamic domain (i.e., any passage or knowledge base that the user provides) in natural language form and the system will generate the most precise answers and provide those answers in natural language form. Along with this, a set of important keywords are also generated that are embedded in the system in order to help the user to understand the context of the passage and ask the questions. A state is reproduced where the user communicates with a person in a particular profession as close as achievable.

It is not possible to generate a standard algorithm which understands a language and reply to questions asked by the user. To alleviate such type of concerns, artificial neural networks was added. ANN is a knowledge processing paradigm which is animated by biological nervous systems. Recurrent neural network is one of the successful techniques of artificial neural network that has applied for text understanding. One of the approaches used by RNNs is the notion they might be capable of relating past knowledge to the existing task. Figure 3 shows an example of keyword generation along with question-answer system.

However, there were some exceptional cases where the system failed to provide results as required as shown in Fig. 4. This happens due to training larger sentences

```
PASSAGE: The Taj Mahal (/tə:dʒ ma'hol, ,tə:ʒ-/;[4] Hindi: ताज महल [ta:dʒ 'ma:h(a)l], meaning "Crown of the Palaces"
mahal
world
emperor
million
admired
gardens
ta:dʒ
architects
architect
marble
project
wall
heritage
construction
jahan
ASK QUESTION : Who built the Taj Mahal?
ANSWER      : Shah Jahan
ASK QUESTION : When was it built?
ANSWER      : 1643
ASK QUESTION : What was the cost of construction?
ANSWER      : 32 million rupees
```

**Fig. 3.** Example 1 (History)

```

PASSAGE: The Central Board of Secondary Education (CBSE) is a national
testing
test
tests
conducts
conducted
conduct
conducting
schools
cbse
examinations
examination
exam
union
entrance
medical
navodaya
research
central
jee
eligibility
ASK QUESTION : When was CBSE established?
ANSWER : 10 November 2017
ASK QUESTION : Who conducts the JEE mains?
ANSWER : National Testing Agency
ASK QUESTION : How many schools in India follow CBSE?
ANSWER : 19,316

```

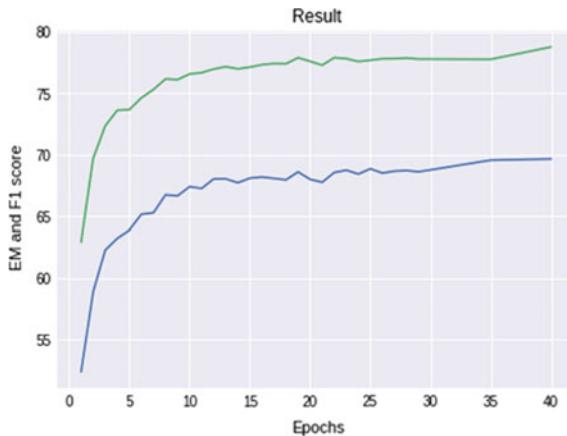
**Fig. 4** Example 2(Education)

or conversations. In that case, the network is much slower to train and the network essentially gives the closest answers (according to the tags present) but it appears to have combined some fundamental semantic (Fig. 5).

One of the competitive machine comprehension benchmarks is SQuAD and the table only contains the best-performing systems. This proposed system (QA with keyword) reaches 70.8% exact match and 79.8% F1 scores on the test set, which exceeds all the proclaimed results as shown in Table 1.

## 5 Future Work

In the current model, the questions are independent of each other. To generate the dependencies among the questions, all the previously asked questions along with the answers should be feed to the encoder. Once the encoder/decoder is trained with such data, it can be used as a dialogue-based system.



**Fig. 5** EM and F1 score

**Table 1** Different EM and F1 score

Method	Dev		Test	
	EM	F1	EM	F1
Dynamic coattention networks	65.4	75.6	66.2	75.9
Multi-perspective matching	66.1	75.8	65.5	75.1
R-Net	N/A	N/A	71.3	79.7
BiDAF	67.7	77.3	68.0	79.7
Our model (QA with keywords)	70.8	79.8	70.0	79.0

## 6 Conclusion

This work proposed an idea of keyword generation in question answering system (QA with keyword). The proposed question answering system was trained using Bidirectional LSTM for passage apprehension and RNN-based encoder for question generation. The self-attention mechanism has also been used here to establish correlation between current words and previous parts of the sentence. Our question answering system is successful in reaching EM and F1 score on Stanford Question Answering Dataset (SQuAD) which are comparable to the state-of-art system (as mentioned in Table 1). Additionally, we believe that our paradigm is conceptually simpler than most of the present systems. Therefore, we conclude the applicability of this research in the area of natural language processing.

## References

1. Bordes A, Usunier N, Chopra S, Weston J (2015) Large-scale simple question answering with memory networks. ArXiv preprint [arXiv:1506.02075](https://arxiv.org/abs/1506.02075)
2. Wang L, Xiong Y, Wang Y, Qiao DL, Tang X, Van Gool L (2016) Temporal segment networks: towards good practices for deep action recognition. In: European conference on computer vision, pp 20–36
3. Saad MS (2019) Adaptive artificial intelligent Q&A platform. ArXiv preprint [arXiv:1902.02162](https://arxiv.org/abs/1902.02162)
4. Chen D, Fisch A, Weston J, Bordes A (2017) Reading wikipedia to answer open-domain questions. ArXiv preprint [arXiv:1704.00051](https://arxiv.org/abs/1704.00051)
5. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data, pp 1247–1250 (2008)
6. Berant J, Chou A, Frostig R, Liang P (2013) Semantic parsing on freebase from question-answer pairs. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1533–1544 (2013)
7. Fader A, Zettlemoyer L, Etzioni O (2014) Open question answering over curated and extracted knowledge bases. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1156–1165
8. Caruana R (1997) Multitask learning. Machine Learning, 28(1):41–75, Springer
9. Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on machine learning, pp 160–167 (2008)
10. Allam A, Haggag M (2012) The question answering systems: a survey. Int J, Res Rev Inf Sci (IJRRIS) 2(3)
11. Bhaskar P, Pakray P, Banerjee S, Banerjee S, Bandyopadhyay S, Gelbukh A (2012) Question answering system for QA4MRE CLEF 2012, CLEF (Online Working Notes/Labs/Workshop)
12. Tao F, Liu G (2018) (Advanced LSTM: a study about better time dependency modeling in emotion recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2906–2910 (2018)
13. Ling W, Luis T, Marujo L, Astudillo R, Amir S, Dyer C, Black A, Trancoso, I (2015) Finding function in formic Compositional character models for open vocabulary word representation. ArXiv preprint [arXiv:1508.02096](https://arxiv.org/abs/1508.02096) (2015)
14. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. ArXiv preprint [arXiv:1409.0473](https://arxiv.org/abs/1409.0473)
15. Hill F, Bordes A, Chopra S, Weston J (2015) The goldilocks principle: Reading children’s books with explicit memory representations. ArXiv preprint [arXiv:1511.02301](https://arxiv.org/abs/1511.02301)
16. Graves A, Wayne G, Danihelka I (2014) Neural turing machines. ArXiv preprint [arXiv:1410.5401](https://arxiv.org/abs/1410.5401)
17. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000 + questions for machine comprehension of text. ArXiv preprint [arXiv:1606.05250](https://arxiv.org/abs/1606.05250)
18. Choi E, Hewlett D, Uszkoreit J, Polosukhin I, Lacoste A, Berant J (2017) Coarse-to-fine question answering for long documents. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 209–220
19. Iyyer M, Manjunatha V, Boyd-Graber J, Daum III H (2015) Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long Papers), vol 1, pp 1681–1691
20. Rocktaschel T, Grefenstette E, Hermann K, Kovciský T, Blunsom P (2015) Reasoning about entailment with neural attention. ArXiv preprint [arXiv:1509.06664](https://arxiv.org/abs/1509.06664)
21. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. ArXiv preprint [arXiv:1406.1078](https://arxiv.org/abs/1406.1078)

22. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv preprint [arXiv:1412.3555](https://arxiv.org/abs/1412.3555)
23. Kadlec R, Schmid M, Bajgar O, Kleindienst J (2016) Text understanding with the attention sum reader network. ArXiv preprint [arXiv:1603.01547](https://arxiv.org/abs/1603.01547)
24. Cui Y, Chen Z, Wei Z, Wang S, Liu T, Hu G (2016) Attention-over-attention neural networks for reading comprehension. ArXiv preprint [arXiv:1607.04423](https://arxiv.org/abs/1607.04423)
25. Sukhbaatar S, Weston J, Fergus R (2015) End-to-end memory networks. *Adv Neural Inf Proc Syst* 2440–2448
26. Gulcehre C, Ahn S, Nallapati R, Zhou B, Bengio Y (2016) Pointing the unknown words. ArXiv preprint [arXiv:1603.08148](https://arxiv.org/abs/1603.08148)
27. Yu Y, Zhang W, Hang C, Xiang B, Zhou B (2015) Empirical study on deep learning models for question answering. ArXiv preprint [arXiv:1510.07526](https://arxiv.org/abs/1510.07526)
28. Richardson M, Burges C, Renshaw E (2013) Mc-test: a challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 193–203
29. Berant J, Srikumar V, Chen P, Vander Linden A, Harding B, Huang B, Clark P, Manning C (2014) Modeling biological processes for reading comprehension. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1499–1510 (2014)
30. Yang Y, Yih W, Meek C (2015) Wikiqa: a challenge dataset for open-domain question answering. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 2013–2018

# Classification of LISS-III Image Using Fuzzy Logic



Anand Upadhyay, Sonam Mishra, and Aishwarya Khavadkar

**Abstract** The LISS-III is the multi-phantom camera working in four groups. The main reason behind accompanying the work is to apply calculation dependent on regulated characterization of systems to comprehend the land spread and land utilized region in Mumbai. Here, we have used the IRS P6 LISS-III satellite picture of Mumbai locale is utilized to group the regions of Mumbai, Navi Mumbai, and Thane district. The classifier utilized is a fuzzy inference system and band pictures. The various regions of Mumbai locale are grouped, for example, zone secured by mangroves, forest, water, and developed area. It is been seen that the accuracy of fuzzy inference system is 77.88%.

**Keywords** Image processing · Fuzzy logic · Fuzzy inference system · Supervised learning

## 1 Introduction

Image classification is a unique among the most significant pieces of image analysis. Two essential methodologies are supervised and unsupervised learning. In two types, the process can be seen as one that determines the set that each pixel has. In the case of directed characterization, the sets are known beforehand but, due to the uncertain order, the sets are ambiguous. The majority of the investigations in order are carried out as a supervised [1]. In the supervised strategies, a model is developed dependent on the cluster known occurrences and will recognize new articles [2].

---

A. Upadhyay (✉) · S. Mishra · A. Khavadkar

Department of Information Technology, Thakur College of Science and Commerce,  
Kandivali (E), Mumbai, Maharashtra 400101, India

e-mail: [anandhari6@gmail.com](mailto:anandhari6@gmail.com)

S. Mishra

e-mail: [sm4940268@gmail.com](mailto:sm4940268@gmail.com)

A. Khavadkar

e-mail: [aishwaryak13@gmail.com](mailto:aishwaryak13@gmail.com)

There is a bottleneck in the supervised group that they tend to focus less on suspicious symptoms because the preparation set covers only a few occasions [3]. Additionally, the preparation dataset created are helpful just when the pictures are concurrent, or for the pictures choose under a similar condition with similar classes. But the fact is that actual land and land utilize are regularly used to the contrary, their real implications are very. Land utilized mapping is different and is the most significant and run of the mill uses of remote detecting information [4]. Land utilized refers to the surface that extends over the land, whether it is vegetation, urban foundations, water, open soil, or others. Identifying, designing, and mapping land cover is important for arranging examinations, assets, boards, and exercises around the world. Recognizable proof of land cover sets up the benchmark from which checking exercises can be performed and gives the ground spread data to gauge topical maps. Land use implies the reason that land serves that is living space, or agribusiness. Land-use applications include benchmark mapping and consequent checking because convenient data is required to realize what current amounts of land are in use and to isolate land-use changes from year to year.

## 2 Fuzzy Logic

In the last few years, fuzzy logic has been used for various domains and problems, but fuzzy logic is a fairly recent theory [5]. The applications are widely used for process control, operational research, management economics, and decision-making. For this paper, we have used fuzzy inference system that formulates the mapping given by input to an output using its technique, points which need to be taken care while implementing fuzzy interface member functions are fuzzy logic operators and if then rules [6]. We have used Mamdani method to implement this technique; this is the most commonly used fuzzy method, and it accepts the output membership function to be fuzzy, once the aggregation is processed, and each output label requires fuzzy sets that require definition [7].

### 2.1 Fuzzy Sets

Fuzzy set is a concept of fuzzy logic [8]. The fuzzy set is a set without a clearly defined boundary. It contains elements with partial values. Fuzzy logic is a form of multi-valued logic in which the true value of the variable inclusive (0 and 1) can be any real number between the two. Fuzzy logic is a way to understand processing dependent on “degree of truth” instead of standard thing “True or False” [9, 10].

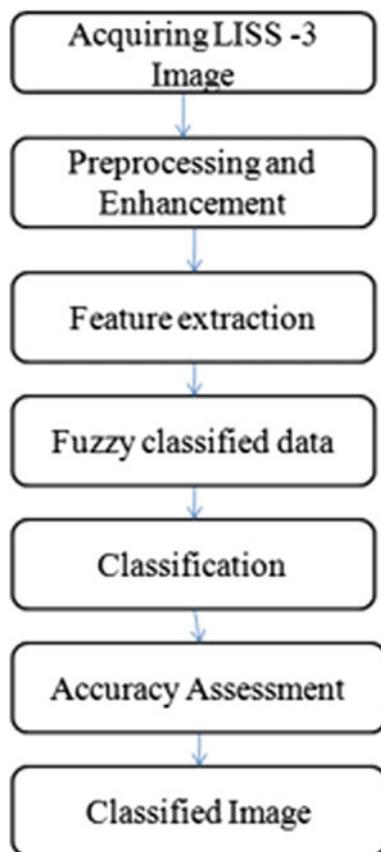
### 3 Methodology

It is the implementation of the proposed algorithm where we apply algorithm and test the data, so the proposed algorithm is implemented using MATLAB simulation toolbox. It classifies the image based on its characteristics. Fuzzy inference systems are used to analyze data and show effects [7].

### 4 Flow Chart

See Fig. 1.

**Fig. 1** Work flow



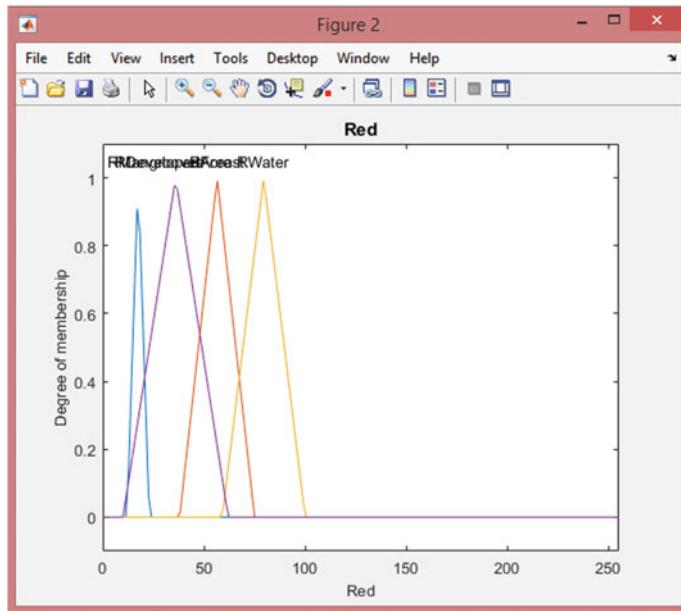
#### 4.1 Fuzzy Classified Data

See Figs. 2, 3, 4 and 5.

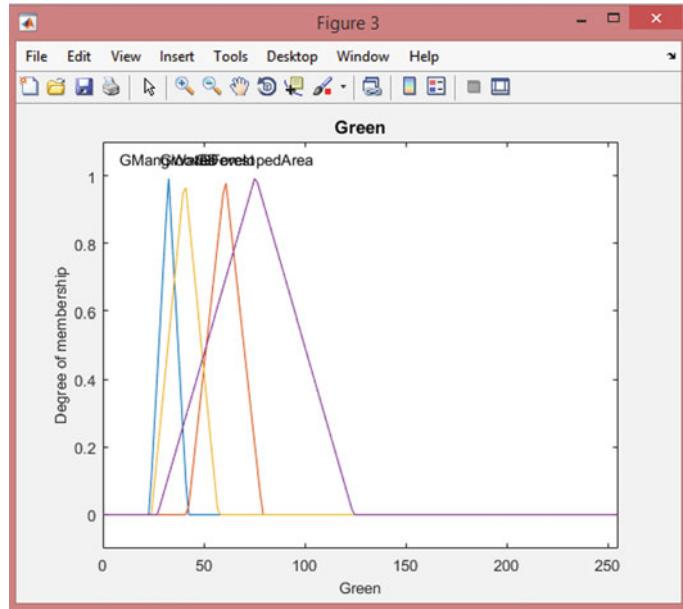
### 5 Result and Observation

The confusion matrix is vital method for testing the efficiency of any classification algorithm. The performance evaluation of any classifier is one of the major factors which suggest the importance of any classifier for any problem solving method [11, 12]. The confusion matrix consists of properly classified and misclassified data where diagonal elements show the classified whereas the non-diagonal elements show misclassified data. The confusion matrix is easy way to calculate accuracy of any classifier. Here, confusion matrix is used to calculate the accuracy of general regression neural network classifier.

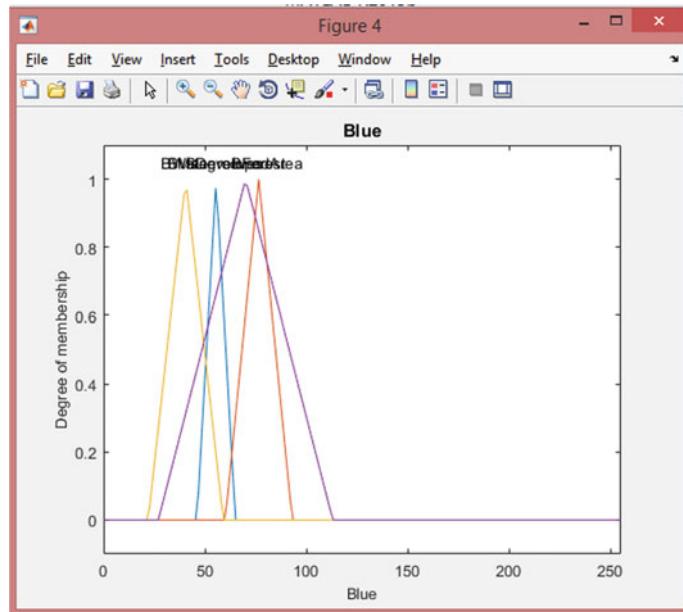
The confusion matrix based accuracy assessment results show that the accuracy of fuzzy classifier for classification of LISS-III satellite image is 77.88% which is good with reference to LISS-III satellite image (Figs. 6 and 7).



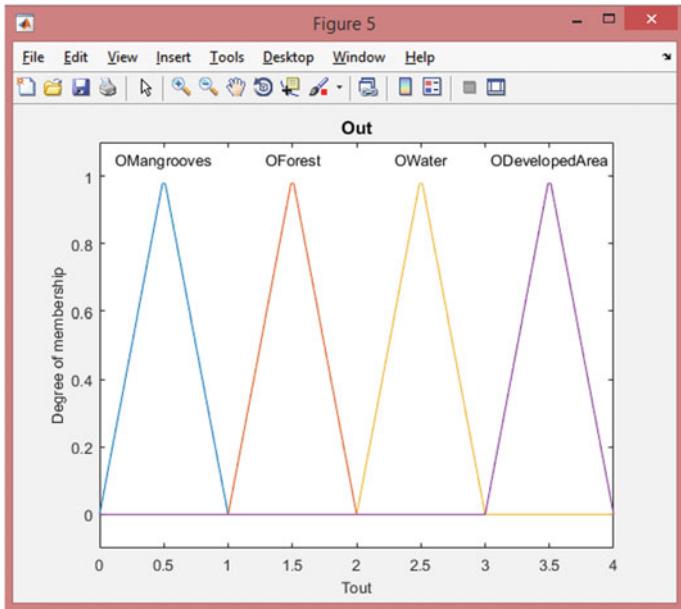
**Fig. 2** Input red for four membership function



**Fig. 3** Input green for four membership function

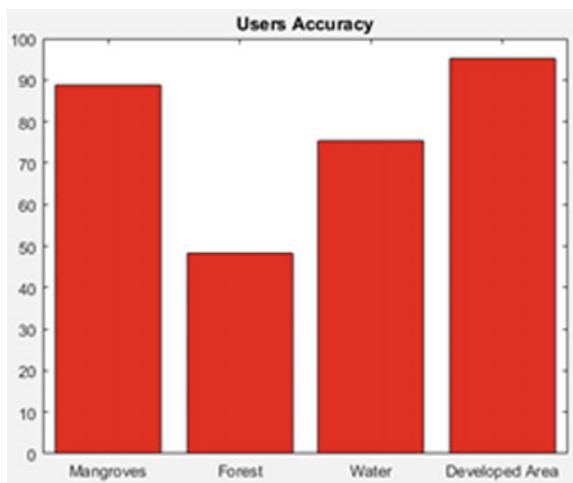


**Fig. 4** Input Blue for four membership function



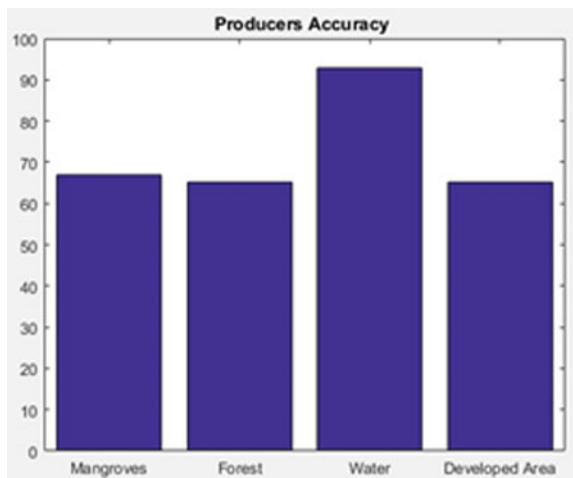
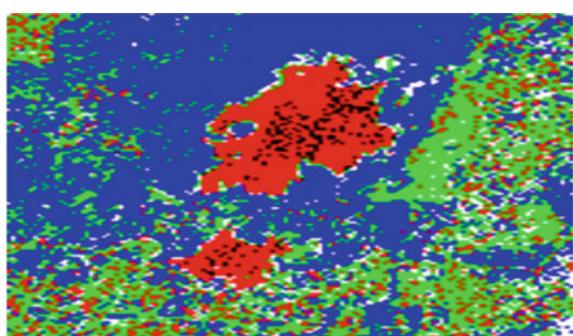
**Fig. 5** output for four different classes

**Fig. 6** Users accuracy



### 5.1 Unclassified and Classified Images

The implementation and testing of the image using classification techniques helps us to classify mangroves, forest, water, and developed area (Figs. 7, 8, 9 and 10; Tables 1 and 2).

**Fig. 7** Producers accuracy**Fig. 8** False color image before classification**Fig. 9** Colored image after classification



**Fig. 10** Color of region's

**Table 1** Accuracy assessment fuzzy inference system

Classes	Mangroves	Forest	Water	Developed area	Total	User accuracy (%)
Mangroves	440	56	0	0	496	87.70
Forest	108	119	17	3	247	48.18
Water	0	2	250	80	332	75.30
Developed area	0	6	2	156	164	95.12
Total	548	183	269	239	965	
Producers accuracy (%)	80.29	65.02	92.94	65.27		77.88

$$\text{Accuracy} = (965/1239)*100 = \mathbf{77.88\%}$$

**Table 2** Overall accuracy

S. No	Classified used	Accuracy
1	Fuzzy inference system	77.88%

## 6 Conclusion

Satellite image-based classification of land use and land cover is a very wide field of study and research, and so many people are researching in terms of efficient algorithms, performance, data handling, training, or time making. So, in the same relation here, the fuzzy inference system method has been used to classify the LISS-III satellite image and the accuracy is calculated using the confusion matrix. The results show that the accuracy of the fuzzy inference system method is 77.88%. This study also shows that with increasing the size of the training set, the classification accuracy increases but to a certain extent. Effective accuracy can also be achieved through increasing the number of training samples and giving a better image. Many training samples depend on the complexity of the study area. If the study area is simple and has well-defined crisp classes, then fewer pixels can also give effective accuracy.

## References

1. Mahashwari T, Asthana A (2013) Image enhancement using fuzzy technique. IJRREST 2(2). ISSN 2278-6643
2. Janhavi Shirke NMS (2016) Multi-label classification of a scene image using fuzzy logic. Int J Comput Appl (0975–8887) Emerg Trends in Comput
3. Younes AA, MSH T, Akdag H (2005) Color image profiling using fuzzy sets. Turk J Elec Engin 13(3)
4. Priyadarshini M, Karthi R, Sangeethaa SN, Premalatha R, Tamilselvan KS (2013) Implementation of fuzzy logic for the high-resolution remote sensing images with improved accuracy. IOSR J Electr Electron Eng (IOSR-JEEE) 5(3):13. e-ISSN: 2278-1676, p-ISSN:2320-3331
5. Souverville S, Rosales JA, Gallegos FJ, Dehesa M, Hernández IV, Lozano LV (2015) Fuzzy logic applied to improvement of image resolution using gaussian membership functions. Res Comput Sci 102:77–88 (rec. 2015-03-28; acc.2015-07-15)
6. Mustafa NBA, Khaleel Ahmed S, Ali Z, Yit WB, Abidin AAZ, Md Sharrif ZA (2009) Agricultural produce sorting and grading using support vector machines and fuzzy logic. In: 2009 IEEE international conference on signal and image processing applications, pp. 391–396
7. Sharma M, Gupta R, Kumar D, Kapoor R (2011) Efficacious approach for satellite image classification. J. Electr Electron Eng Res 3(8):143–150. ISSN: 2141–2367 (©2011 Academic Journals)
8. Park V, Lee H-K (1998) Fuzzy logic based satellite image classification: generation of fuzzy membership function and rule from training set. IAPR workshop on machine vision applications. Nov. 17–19. 1998, Makuhari, Chiba Japan
9. Kamra A, Rani K (2012) An improved method for image enhancement using fuzzy approach. IRACST (IJCSITS) 2(6). ISSN: 2249–9555
10. Kaur A, Kaur A (2012) Comparison of mamdani-type and sugeno-type fuzzy inference systems for air conditioning system. IJSCE 2(2)
11. Shenbagavalli R, Ramar K (2013) Satellite image edge detection using fuzzy logic. Int J Eng Sci (IJES) 2(1): 47–52. ISSN: 2319–1813 ISBN: 2319–1805.
12. Naganur, HG, Sannakki HG, Rajpurohit, VS Arunkumar R (2012) Fruits sorting and grading using fuzzy logic. Int J Adv Res Comput Eng Technol (IJARCET) 1(6). ISSN: 2278--1323

# Optimized Text Classification Using Deep Learning



Neeti Sangwan and Vishal Bhatnagar

**Abstract** As there is tremendous hike in the amount of data created in the world, the need for text classification is on rise. Data from all the online sources: e-mails, web pages, social media, chats, and more results in a huge amount of unstructured text. To extract the information from the text that is unstructured in nature is very cumbersome and time-taking. Therefore, text classification becomes a pre-requisite for the businesses to improve the decision-making process. Different deep learning-based models for text classification with respect to different activation functions are analyzed in the paper.

**Keywords** Classification · Deep learning · Activation function · Convolution

## 1 Introduction

Text classification deals with classifying the unlabeled text document to their corresponding category. There are many issues that exist with big data. With the growing data, the prevalent issue comes out is its unstructured nature.

Large amount of data reside in the freeform that may contain very relevant information but as it is not suitable for the traditional models for analysis that relevant information may be lost. This issue is of concern for both the industries as well as for consumers. Industries fail to keep up with it. There is large amount of information exists within the big data but it is unverified data. Due to unstructured data, some analysis may outcome in inaccurate results that lead to erroneous decision making and makes the process of extraction of relevant information very costly.

---

N. Sangwan (✉)

GGS Indraprastha University, New Delhi, India

e-mail: [neetisangwan@gmail.com](mailto:neetisangwan@gmail.com)

MSIT, New Delhi, India

V. Bhatnagar

Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, India

e-mail: [vishalbhatnagar@yahoo.com](mailto:vishalbhatnagar@yahoo.com)

Deep learning improves the processing ability of the systems with low hardware cost. Deep learning provides the revolution to machine learning. Different deep learning networks can be used for text classification. Deep learning networks include Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), long short-term memory (LSTM), and deep belief network (DBN).

In this paper, distinct models based on deep learning are implemented. Different models are evaluated using different activation functions. Activation function decides which neuron is to be activated by using their weighted sum and bias. It performs non-linear transformations to the inputs to make the classifiers learn better and to perform more complex tasks. Different activation functions can be applied according to the requirements. Various activation functions include: Step function is non-differentiable at zero and not capable of updating the weights. To overcome the drawbacks of the step function, sigmoid function comes into role. It signifies the approximate behavior of the dependent variable. It is represented as S-shaped curve. It provides non-linearity present in the data that is essential for accurate results. Because of its differentiable nature, it is suitable for back propagation and gradient descent approaches. Softmax function is a type of sigmoid function, mainly used for the classification-related problems. It is utilized in the output layer of the classifier. Rectified Linear Unit (ReLU) is usually used for deep learning models. It is significantly utilized in the hidden layers of the neural network. It learns faster than the sigmoid activation function.

Various architectures are combined and discussed according to their accuracy and efficiency to find the best one out of them. Section 2 throws some light on the relevant work done by the distinct authors in the history related to the text classification. Section 3 provides the different system architectures that are implemented to perform the classification task efficiently and accurately. Section 4 provides the results corresponding to the implemented architectures and also discusses the performance of architecture with respect to other implemented architectures. Finally, paper is concluded along with some future directions.

## 2 Literature Review

In this section, some light has been thrown on the relevant work in the field of text classification.

Al-Anzi and AbuZeina [1] described the cosine similarity measure as a distance measure to evaluate the text classification of some Arabic text. Singular Value Decomposition (SVD) is exploited by the authors to fetch out the text-based features with the help of latent semantic indexing. Authors explored the various text mining techniques and their application domains [2]. Sachan et al. [3] explained a deep learning model: bidirectional LSTM network for text classification with both the supervised as well as unsupervised approaches. In [4], a strategy for classifying sentiments with the usage of Latent Semantic Indexing (LSI) is demonstrated. Main motive of using LSI is to grade documents with respect to a given query. A mechanism was provided

to generate positive and negative queries automatically. These queries were then used to obtain negative and positive scores so that a decision could be made on the basis of these scores. This method was not only aimed at separating the positive and negative reviews, but also at providing ranked lists of positive or negative comments. These lists are very important to carry out significant reviews from the top of the negative list, and shining reviews from the top of the positive list.

In [5], the authors utilized 2D-convolution operation to get the meaningful features of the text. LSTM-based integrated architecture is proposed that seizes long-term sentence dependencies. In [6], three distinct methods are proposed for text modeling based on information sharing. An adversarial framework is proposed for multi-task learning to find out general task-independent features [7]. Proposed approach overcome the issues of noise from other tasks and task-specific features in shared features by providing the shared and non-shared feature space that are independent from each other. Yang et al. A stratified network for classifying the text document is introduced [8]. With hierarchical structure model contains two levels of attention that is implemented at word as well as sentence level to deal with significant and non-significant data differentially. A novel model: Very Deep Convolutional Neural Network (VDCNN) is introduced for text processing that works at the atomic level i.e., character level of the text [9]. Lee and Demoncourt [10] introduced an approach build on CNN and RNN for classification of sequential short text. A novel approach build on DBN is proposed to rectify the issue of sparse high dimensional matrix computation of text [11]. In [12], a joint CNN and RNN framework is proposed to reduce the required number of convolution layers for long-term dependencies. CNN-based architecture using dynamic K-max pooling is designed to frame the sentences in the document semantically [13]. In [14], various techniques and issues related to the textual data categorization is addressed. In [15], classification using deep neural network is discussed to find the attributes. Semantics of the medical data is represented and classified with the help of multiple layers of CNNs [16]. In [17], a novel back off model for better classification. In [18], support vector decomposition is implemented on local area under each class to improve the classification process.

### 3 System Architecture and Methodology

Different neural networks are implemented to classify the text automatically and efficiently. These neural networks use the layered architecture that consists of different layers with number of activation functions. Architectures that are implemented with different activation functions are discussed below.

IPython console		
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 150)	0
embedding_1 (Embedding)	(None, 150, 50)	50000
conv1d_1 (Conv1D)	(None, 146, 128)	32128
max_pooling1d_1 (MaxPooling1D)	(None, 29, 128)	0
conv1d_2 (Conv1D)	(None, 25, 128)	82048
max_pooling1d_2 (MaxPooling1D)	(None, 5, 128)	0
flatten_1 (Flatten)	(None, 640)	0
dense_1 (Dense)	(None, 128)	82048
dense_2 (Dense)	(None, 1)	129
<hr/>		
Total params:	246,353	
Trainable params:	246,353	
Non-trainable params:	0	

**Fig. 1** CNN layered architecture

### 3.1 Convolutional Neural Network (CNN)

This architecture is based on general neural network with weights and biases for learning. It gets input in the input layer and passes it through distinct hidden layers followed by output layer. In this architecture, convolution layer, meant for computations, forms the significant component of the Convolutional Neural Network. Max-pooling layer is meant for lessening the requirement of the limiting factors and processing in the network using max function. This architecture utilized two activation functions: Sigmoid and ReLU to increase the performance. Figure 1 depicts the implementation details for CNN architecture.

### 3.2 Recurrent Neural Network-Long Short-Term Memory 1 (RNN-LSTM 1)

This class of artificial neural network is suitable for the time-based processes. It contains its internal state (memory) to work on the series of the inputs. It usually has the short-term memory, that is why RNN uses LSTM as one of the layer. LSTMs are kind of RNN meant for long-term learning. The combination of LSTM helps RNN

to remember something across a sentence. It uses the single activation function i.e., Softmax for activating the neurons in the network.

### 3.3 Recurrent Neural Network-Long Short-Term Memory 2 (RNN-LSTM2)

This architecture is built on the RNN that have internal remembering unit along with LSTM to prolong the remembering power of the network to get the more efficient results. To provide the capability of differentiation, sigmoid activation function is utilized.

### 3.4 Recurrent Neural Network-Long Short-Term Memory 3 (RNN-LSTM3)

In this architecture, RNN is implemented along with the LSTM layer to strengthen the memory of the neural network. ReLU is the activation function that has been utilized in the architecture for choosing the neuron to be activated using the weight and bias. Figure 2 shows the RNN-LSTM architecture that uses a single activation function for the activation of the neurons.

IPython console		
Layer (type)	Output Shape	Param #
inputs (InputLayer)	(None, 150)	0
embedding_2 (Embedding)	(None, 150, 50)	50000
lstm_1 (LSTM)	(None, 64)	29440
dense_3 (Dense)	(None, 1)	65

Fig. 2 LSTM with one activation function

IPython console		
Layer (type)	Output Shape	Param #
inputs (InputLayer)	(None, 150)	0
embedding_3 (Embedding)	(None, 150, 50)	50000
lstm_2 (LSTM)	(None, 64)	29440
FC1 (Dense)	(None, 256)	16640
activation_1 (Activation)	(None, 256)	0
dropout_1 (Dropout)	(None, 256)	0
out_layer (Dense)	(None, 1)	257
activation_2 (Activation)	(None, 1)	0

**Fig. 3** RNN with LSTM layer architecture and two activation functions

### 3.5 Recurrent Neural Network-Long Short-Term Memory 4 (RNN-LSTM4)

In this architecture, RNN is combined with LSTM and two kinds of activation functions. These activation functions increase the overall accuracy of the approach. Figure 3 depicts the RNN-LSTM framework along with the utilization of activation functions: ReLU and Sigmoid for the activation purpose.

### 3.6 Recurrent Neural Network-Bidirectional Long Short-Term Memory (RNN-Bi-LSTM)

In this framework of RNN and bidirectional LSTM, the signal in the neural network propagates in both the directions: forward as well as backward direction at a same time. In unidirectional LSTM, the inputs are only from the past, so, the information in the past is only remembered in the network. Therefore, bidirectional LSTM is used to get the inputs from past to future when LSTM runs forward and from future to past when LSTM runs backward. RNN with bidirectional LSTM and two activation functions: Sigmoid and ReLU is implemented in the architecture.

All the above described deep learning-based models are implemented and evaluated. These models are trained and validated on two well-known datasets:

**Dataset 1:** It consists of 5574 ham-spam messages. 87% of the messages considered are ham messages and remaining 13% messages are spam. Out of 5574 messages, 4459 messages are considered for the training of the models and 1115 messages are considered for the validation of the models.

**Dataset 2:** It consists of 7540 product reviews of two classes from Amazon. Out of which, 6032 are taken for the training and 1508 reviews are considered for the validation purpose.

## 4 Results and Discussion

Accuracy of the distinct deep learning architectures for classifying text is delineated in Table 1. Increased accuracy can be achieved with different algorithms.

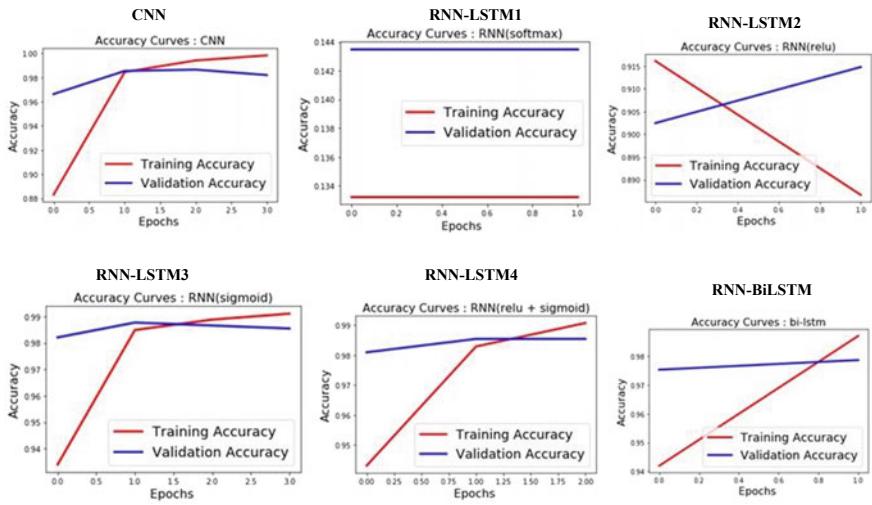
Figures 4 and 5 demonstrate the training and validation accuracy of all the models implemented on the dataset 1 and dataset 2, respectively. Graphs show the accuracy achieved by all the models that are using different activation functions and layers arrangements with the increasing epochs.

From the above experimentation and the result plots, following key points are implied:

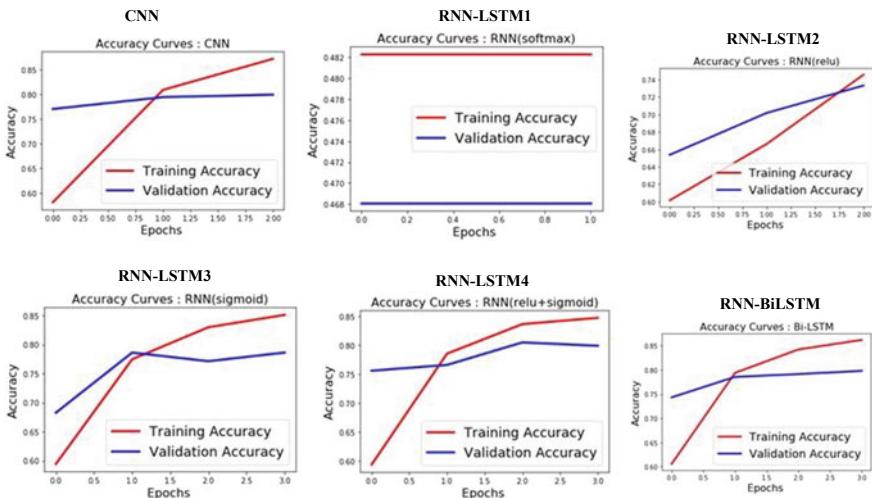
- CNN has achieved good validation accuracy with high consistency, also RNN has achieved high accuracy but they are not that consistent throughout all the datasets.
- Softmax activation function was found to be the worst architecture to implement for production ready scenarios.
- CNN model has outperformed the other two models (RNN and HAN) in terms of training time; however, Bi-LSTM can perform better than CNN and RNN(LSTM) if there is huge dataset.

**Table 1** Overall result comparison

Dataset	Methodology	Algorithm	Activation function	Accuracy
Dataset 1	CNN	CNN	ReLU + Sigmoid	98.3
	RNN-LSTM1	RNN-LSTM	Softmax	15.2
	RNN-LSTM2	RNN-LSTM	ReLU	91.1
	RNN-LSTM3	RNN-LSTM	Sigmoid	97.9
	RNN-LSTM4	RNN-LSTM	ReLU + Sigmoid	98.6
	RNN-BI-LSTM	RNN-BI-LSTM	ReLU + Sigmoid	98.1
Dataset 2	CNN	CNN	ReLU + Sigmoid	80.2
	RNN-LSTM1	RNN-LSTM	Softmax	47.1
	RNN-LSTM2	RNN-LSTM	ReLU	72.8
	RNN-LSTM3	RNN-LSTM	Sigmoid	77.3
	RNN-LSTM4	RNN-LSTM	ReLU + Sigmoid	79.0
	RNN-BI-LSTM	RNN-BI-LSTM	ReLU + Sigmoid	78.8



**Fig. 4** Training accuracy and validation accuracy of the models on dataset 1



**Fig. 5** Training accuracy and validation accuracy of the models on dataset 2

- For dataset 1, the huge difference between the two categories has made it easy for every model to come out with high accuracy every time.
- For dataset 2, the little difference between two categories requires more training with good model. In that scenario also, CNN had outperformed the other models.
- Overall, CNN and LSTM combined with Relu and sigmoid are doing good on text classification as compared to other models and variants. Softmax function performs worst among all the activation functions.

## 5 Conclusions

In this paper, different deep learning-based algorithms are presented and compared for their performances in text classification. In this analysis, different activation functions and their combinations are utilized to find the best combination for good text classification. This paper addresses the deep learning models and significant improvement is achieved in classification accuracy, hardware cost, and processing time. Deep learning-based classification of text ease the process of decision making. Result shows that LSTM network along with ReLu and sigmoid activation functions classifies the text with highest accuracy. More attention is required for Bi-LSTM network for better classification of text. Enhancement in the model helps in improving the performance and it will be very useful for the various applications related to classification system.

## References

1. Al-Anzi FS, AbuZeina D (2017) Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing. *J King Saud Univ-Comput Inf Sciences* 29(2):189–195
2. Allahyari M, Pouriyeh S, Assefi M., Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) A brief survey of text mining: Classification, clustering and extraction techniques. ArXiv preprint [arXiv:1707.02919](https://arxiv.org/abs/1707.02919)
3. Sachan DS, Zaheer M, Salakhutdinov R (2018) Revisiting LSTM networks for semi-supervised text classification via mixed objective function. *Proc AAAI Conf on Artif Intell* 33:6940–6948
4. Saqib SM, Kundi FM, Ahmad S (2018) Unsupervised learning method for sorting positive and negative reviews using LSI (latent semantic indexing) with automatic generated queries. *Int J Comput Sci Network Secur* 18(1):56–62
5. Zhou P, Qi Z, Zheng S, Xu J, Bao H, Xu B (2016) Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. ArXiv preprint [arXiv:1611.06639](https://arxiv.org/abs/1611.06639)
6. Liu P, Qiu X, Huang X (2016) Recurrent neural network for text classification with multi-task learning. ArXiv preprint [arXiv:1605.05101](https://arxiv.org/abs/1605.05101)
7. Liu P, Qiu X, Huang X (2017) Adversarial multi-task learning for text classification. ArXiv preprint [arXiv:1704.05742](https://arxiv.org/abs/1704.05742)
8. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480–1489
9. Conneau A, Schwenk H, Barrault L, Lecun Y (2016) Very deep convolutional networks for text classification. ArXiv preprint [arXiv:1606.01781](https://arxiv.org/abs/1606.01781)
10. Lee JY, Dernoncourt F (2016) Sequential short-text classification with recurrent and convolutional neural networks. ArXiv preprint [arXiv:1603.03827](https://arxiv.org/abs/1603.03827)
11. Jiang M, Liang Y, Feng X, Fan X, Pei Z, Xue Y, Guan R (2018) Text classification based on deep belief network and softmax regression. *Neural Comput Appl* 29(1):61–70
12. Hassan A, Mahmood A (2018) Convolutional recurrent deep learning model for sentence classification. *Ieee Access* 6:13949–13957
13. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. ArXiv preprint [arXiv:1404.2188](https://arxiv.org/abs/1404.2188)
14. Sebastiani F (2002) Machine learning in automated text categorization. *ACM Comput Surveys (CSUR)* 34(1):1–47

15. Zeng D, Liu K, Lai S, Zhou G, Zhao J (2014) Relation classification via convolutional deep neural network. In: Proceedings of the 25th international conference on computational linguistics: technical papers, pp 2335–2344 (2014)
16. Hughes M, Li I, Kotoulas S, Suzumura T (2017) Medical text classification using convolutional neural networks. *Stud Health Technol Inform* 235:246–250
17. Nguyen TH, Shirai K (2013) Text classification of technical papers based on text segmentation. In: International conference on application of natural language to information systems. Springer, Berlin, Heidelberg, pp 278–284
18. Liu T, Chen Z, Zhang B, Ma WY, Wu G (2004) Improving text classification using local latent semantic indexing. In: Fourth IEEE international conference on data mining (ICDM'04), IEEE, pp 162–169

# Digital Learning: A Proficient Digital Learning Technology Beyond to Classroom and Traditional Learning



Sanjay Tejasvee, Devendra Gahlot, Rakesh Poonia, and Manoj Kuri

**Abstract** Education is a vital basis of any country towards the development and sustainability of continuous growth for a long term. The system of education should be ahead with new technologies to achieve more benefits at every stage of improvement or enlargement of a nation. Today's era is fully digital era in every sectors; however, the education system is always trying to function according to latest trends, move parallel with new technological aspects throughout the earlier few years. Digital learning and ICT (information and communication technology) tools bond the teachers, academics, parents, experts and institutions. Digital learning also motivates to use latest technologies to deliver, communicate and share to each role players and vice versa. The key concern of the paper is to represent a clear glance of digital learning phenomena and to touch almost every aspects concern to digital learning in brief. This paper will try to deliver from general idea aspect to analysis aspect about digital learning.

**Keywords** Education · Sustainability · Technologies · Enlargement · Digital learning · Parallel · ICT

---

S. Tejasvee (✉) · D. Gahlot · R. Poonia

MCA Department, Government Engineering College Bikaner, Bikaner, Rajasthan 334004, India  
e-mail: [drsanjaytejasvee@gmail.com](mailto:drsanjaytejasvee@gmail.com)

D. Gahlot

e-mail: [dr.devendragahlot@gmail.com](mailto:dr.devendragahlot@gmail.com)

R. Poonia

e-mail: [rakesh.ecb98@gmail.com](mailto:rakesh.ecb98@gmail.com)

M. Kuri

Department of Electronics and Communication, Government Engineering College Bikaner,  
Bikaner, Rajasthan 334004, India  
e-mail: [kuri.manoj@gmail.com](mailto:kuri.manoj@gmail.com)

## 1 Introduction

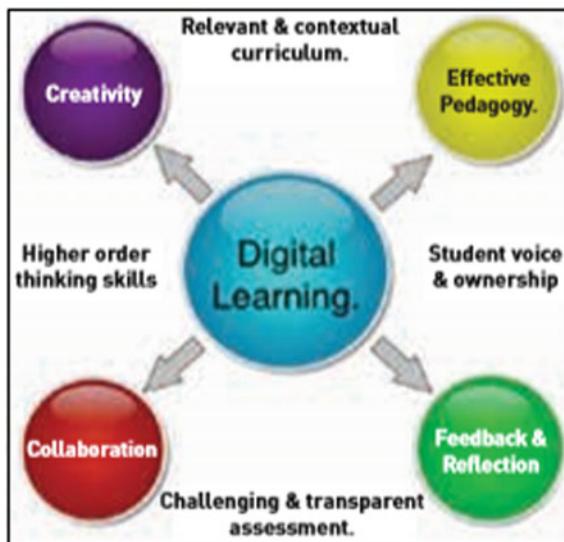
The word ‘digital’ itself is carrying the huge potentiality of various proficient prospects in current era. Subsequent to a high escalation and utilization of Internet, our expectations also lifted high due to digital communication. Today, a normal human being wants more and more possible digitalization in almost every field because he/she very well knew its significance and vast benefits. Technology has turned into an indispensable and vital component of everyone’s life due to almost every task is associated with technology in some way. Digitalization has been impacted with a great force on everything in today’s activities of human at every sector. One of the most considerable fields of digitalization is digital learning in current environment. Most of the people have confusion that digital learning is learning technology but digital learning is learning by or using technology.

Today’s students or learners are frequently known as ‘digital-age-learners’. They can access and share several resources and knowledge by using iPods, computers, smartphones, social media pages, etc., and can get Internet-based skills with collective data and information at every level [1].

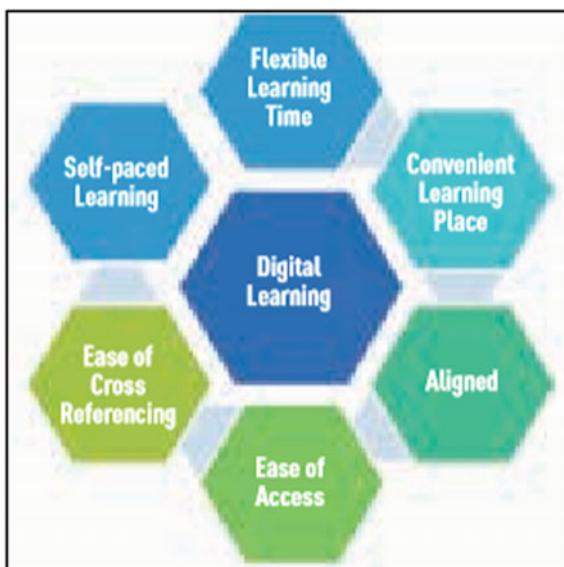
An instructional practice that efficiently uses technological aspects to reinforce of learning experience of learner, called digital learning. It includes a broad range of practices and tools such as online formative assessment, quality improvement of teaching resources with in time, availability of courses, and serving online educational tools and applications in the classroom and within the range of premises, contribution in skilled group of people of practice [2] (Figs. 1, 2 and 3).



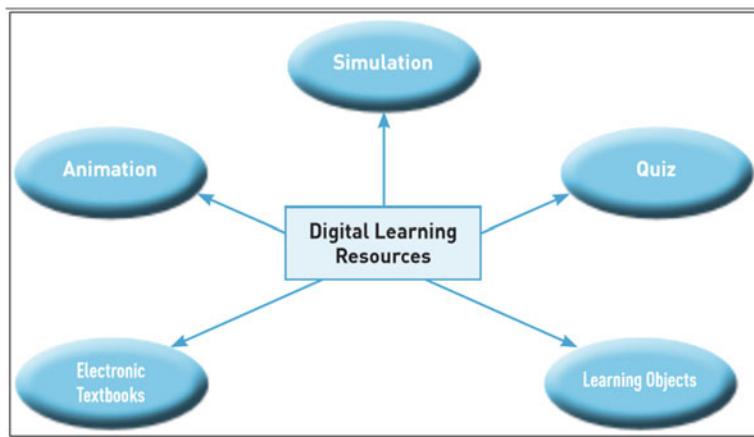
**Fig. 1** ‘Digital Learning’



**Fig. 2** Overview of Digital Learning



**Fig. 3** Features of Digital Learning



**Fig. 4** Types of Digital Learning Resources

## 2 Digital Learning Resources

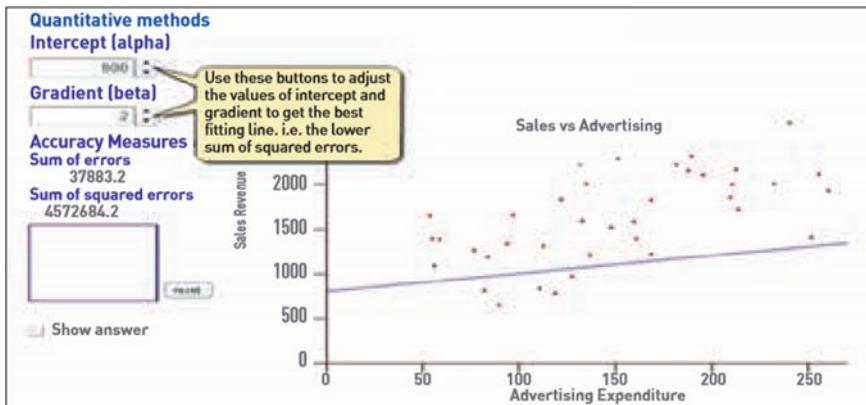
The word ‘digital learning resource’ refers to resources consisted in the framework of a course that support the student’s attainment of their learning goals. These learning resources also used for improving the quality of teaching techniques to better level. Digital learning resources are divided into three vital categories such as the following:

### 2.1 *Simulation*

Simulation assists teachers for explanation of instructional practices dynamically and helps students to check and test their ideas without going through the actual experimentation. An illustration of quantitative method interrupt (alpha) simulation is showing in Fig. 4.

### 2.2 *Animation*

It is a process which is used to generate digitalize images with animation features. The more common name CGI (computer-generated imagery) includes both static and dynamic images, although computer animation only refers to moving images. Recent computer animation frequently uses 3D graphics, low bandwidth and quicker real-time depiction. In animations, learners or students can understand the concept of topic step by step as depicted in Fig. 5.



**Fig. 5** Simulation illustration

### 2.3 Quiz

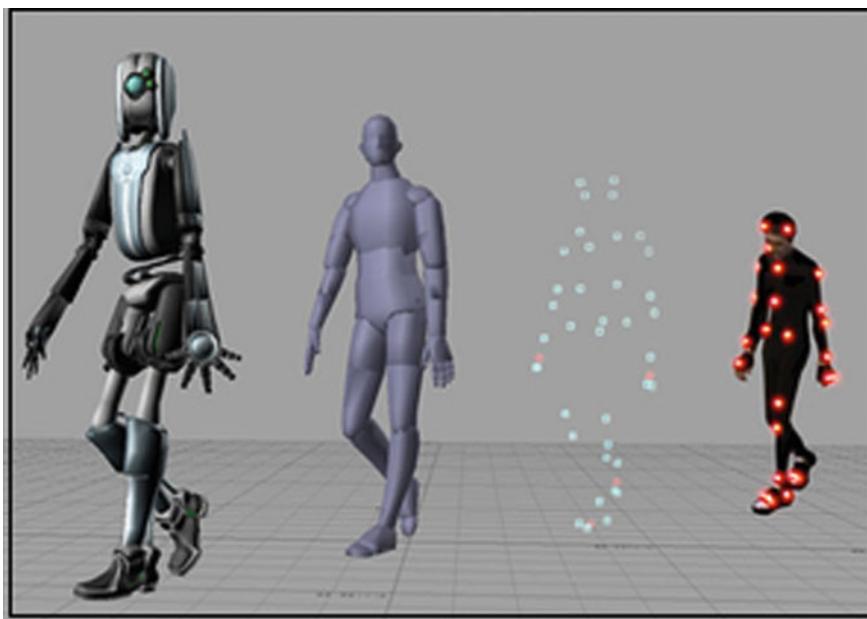
Quiz assists the various students to check their understanding ability about the concepts. Quiz helps learners to increase visibility of the ratio of real accepting and understanding by them. It is just similar to a game to test knowledge or skill related to a particular subject. An illustration of quiz for checking learning and getting feedback is showing in Fig. 6.

### 2.4 Electronic Textbook

It is also known as e-textbooks or digital textbook. It is digitalized publication consisting text, images or combination of both. It can be read on a proprietary digital device (an e-reader) or on a computer that requires special kind of software [3]. In education field, digitalization offers several kinds of books without print on paper at reasonable or less price. In South Korea, all the textbooks were digitalized in 2015 [4].

### 2.5 Learning Objects

Learning object resources are known as digital entities that are basically used for training and education [5]. Digital learning objects are very tiny, modular, discrete units of learning formed for electronic delivery and use. It has its origins in a modular approach to reusable digital instructional materials [6]. Learning objects are



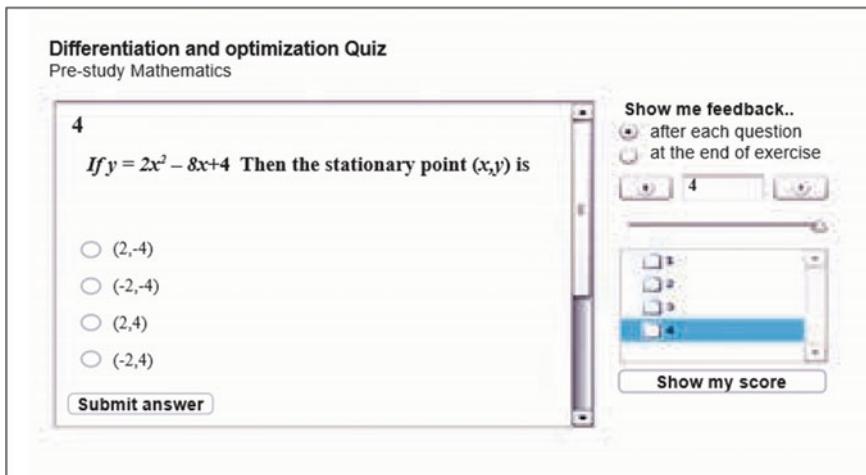
**Fig. 6** Animation illustration ('motion capturing' )

fundamentally an innovative appearance of learning material where content is self-contends, self-controlled and reusable and can be easily cumulative with metadata [7].

### 3 Traditional Versus Digital Learning System

Digital learning put more emphasis on various factors which are not counted as major concern in traditional learning system. Actually, digital learning is a transformation of learning in modern manner towards learners and students. Figure 7 views a lot of factors, out of them mostly not present in traditional learning but in digital learning.

Digital learning provides a special way of learning in its place of boring classroom-based learning [8, 9]. In digital learning, the learners are habitually attached with outside world. As per following, Table 1 shows many differences between traditional and digital learning (Fig. 8).



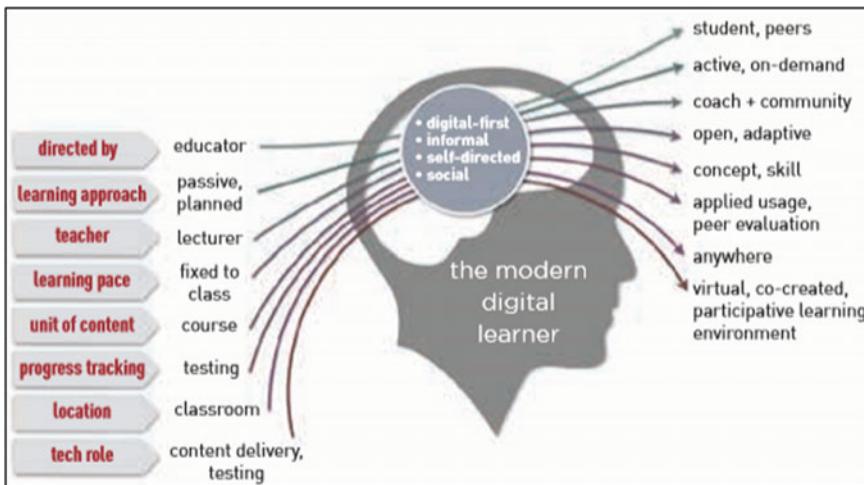
**Fig. 7.** Quiz illustration

**Table 1.** Digital Learning versus Traditional Learning [10]

Learning Parameter	Digital Learning	Traditional learning
Personalized Content	Learning content can be personalized based on the learner preferences	Common content for all learners; personalization is not feasible
Flexi-time	Learn at Any time based on learner convenience	Learners have to make themselves available at specified time
Structure of Information	Information is organized well for cross reference and supported by nice, presentable views	Information is available but it can be time consuming to search for reference and presentation format is rigid.
Learner involvement	Increased involvement of the learners. They have to read, search, re-read, pause and take intermediate learning tests.	Trainers have a bigger responsibility in scheduling the learning, executing it and ensuring that learners have got what they were supposed to from the session.
Training Room	Physical presence of the Learners is not required	Trainers and learners have to be collocated.
Learning Metrics	Data about learning can be easily gathered and analyzed. Important insights can be derived to optimize and improve the learning material	Surveys has to be manually administered to assess training effectiveness. It is not accurate and it is tough to derive insights from the same.

#### 4 Plus (+) Points of Digital Learning

- Digital learning can be possible from anywhere, and it is extremely flexible in nature. Learner can learn digitally at any time which is suitable for learner so he/she can also do different kind of job [11].



**Fig. 8.** Digital transformation of learning (the modern digital learner)

- Digital learning enlarges the engagement and improves the concentration, attention and focus of the learner with the several offered strategies like gaming *f*.
- Digital learning is a money-saving and time-saving method to learn effectively. In digital learning, some courses have very small fees or no fee.
- Digital learning system provides enough time to improve students. It helps learners to improve on their own growth/understanding rate *f*.
- Digital learning offers large, deep and updated recent contents to study, which supports learners to remain up-to-date to most current ideas and technologies.
- Digital learning offers a huge big platform or a framework to the learners where learners can share, communicate, convey, discuss and solve problems and difficulties and also publish their work at distributed level.
- Digital learning provides ownership to the students/learners what to learn and how to demonstrate one's learning.
- Anyone could easily get any information which he/she is searching even if he/she has no idea regarding what he/she is dealing with.
- Without buying books, a learner can get/able to read books and conclude for final solution *f*.
- Learner can obtain training form private tutor. It will give personal focus/attention that a lot are seeking.
- Learners can take assessments periodically to test how learners have understood the concept and can read any books online to improve their skills up to satisfaction of learner.
- Digital learning is more suitable and convenient to learn any concept with the manner by learner choice.

## 5 Minus (–) Points of Digital Learning

- There is a need of proper infrastructure for digital learning.
- Person needs digital devices like smart mobile phone/laptop with Internet, i.e. broadband or mobile data [12].
- Digital learning has interpretability issue. Learner may find different learning material on different apps or platforms but there can be a problem when he/she wants to interact all the stuff at the same time *f*.
- If learner is not dedicated and fully motivated towards his/her goal, then one can easily get distracted in digital learning.
- As Mortaza [13] said, the students' health is affected in many ways which is an essential factor to consider if learners do more works online *f*.
- Students cannot acquire or make a full contact with mentors or teachers. This shortens the motivation and advises that are given by experienced teachers *f*. Cost is one more problem that people face.

## 6 Few Frequent Websites of Digital Learning

There are a large number of websites and apps that deal with which digital learning to help learners to obtain smarter and intellectual out of them few are [14]:

- **Byjus App:** This app teaches concepts in a very easy manner [15].
- **meritnation.com:** This app helps all the students to be skilled in all the subjects and give the answers for all NCERT questions [16].
- **UDEMY:** This app assists learners to study almost everything in technique [17].
- **Solearn App:** This app helps in all the developing programming languages [17]
- **NPTEL:** The courses are offered by IIT and IIS. It motivates learners for joining online classes/courses related to subject or not related to subject which also add value in education [18].

## 7 Conclusion

As the Internet users increased in numbers, many people can obtain benefits of education-related latest things. Digital learning is the prospect of innovative leaning. Demand for digital learning will certainly increase. Today, learners, students, teachers or anyone can catch an answer to their problem which is few clicks away. Digital learning creates a progressive environment. Yes, it is true that if digital learning used wisely, one can change the world because it is very powerful tool to learn huge things through entire the world. Digital learning will help to connect academics and families with fast ease. Digital learning permits a great communication, better information sharing system and improved strong visualization of aim. It also helps to improve

learner's learning capabilities within the system. At last, we can say digital learning is a great step of logical risk and enhancement for the use of recent technology.

## References

1. Burkholder K (2012) The impact of a technology integration academy on instructional technology integration in a Texas School District. Unpublished dissertation. Nova Southeastern University. Ft. Lauderdale, Florida
2. <http://www.digitallearningday.org/>
3. <https://www.techopedia.com/definition/2193/electronic-book>
4. In South Korea, all textbooks will be e-books by 2015. Christian Science Monitor
5. Draft Standard for Learning Object Metadata. IEEE Standard 1484.12.1. Institute of Electrical and Electronics Engineers, New York
6. <https://library.csun.edu/docs/ScholarWorks/LearningObjectsClarification.pdf>
7. Wiley DA (2002) Connecting learning objects to instructional design theory: a definition, a metaphor, and a taxonomy. The instructional use of learning objects. Agency for Instructional Technology, Bloomington, IN. <http://reusability.org/read/chapters/wiley.doc>
8. <https://www.insitu.digital/6-reasonswhy-you-should-consider-usingdigital-learning>
9. <https://www.eztalks.com/elearning/traditional-learning-vs-e-learningwhat-are-their-differences.htm>
10. Anitha P (2018) (Asst. Profesor, Dept of CSE, Presidency Univ., Karnataka) 'Digital learning: a paradigm shift in education' CSI communications-July-2018, p 18. [www.csi-india.org](http://www.csi-india.org)
11. Heider JS (2015) Using digital learning solutions to address higher education's greatest challenges. Publ Res Q 31(3):183–189
12. eLearning in India: advantages and disadvantages. eLearning Industry, 03 Apr 2019. [Online]. Available: <https://eduxpert.in/online-education-india>
13. Mokhtari M (2013) Research on negativity effect on E-Learning. Int J Mobile Netw Commun Telematics, April 2013
14. Revathi AR, Shwetta M (2019) Digital Learning. CSI-Commun 43(4): 21–23
15. Vatsalaya P (2019) Digital Learning: an overview. CSI CommuniCationS, pp 6–7, July 2019
16. Sharma D, Sharma P (2019) Digital Learning: a mysterious potential, vol 43, issue 4, pp 11–13, July 2019
17. Anitha P (2019) Digital Learning—a paradigm shift in education, vol 43, issue 4, pp 18–20, July 2019. [www.csi-india.org](http://www.csi-india.org)
18. Rashid M, Gujri HSM (2019) Introduction and initiatives of Digital Learning schemes in Indian Higher Education, Unnati Gulaty et al, vol 43, issue 4, pp 18–20, July 2019

# Data Security & Future Issues for Cloud Computing



Devendra Gahlot, Sanjay Tejasvee, Kunal Bhushan Ranga,  
and Rishi Raj Vyas

**Abstract** In current vogue, it is required more storage space as well as security being increasing internet user's day by day. Cloud computing is appropriate platform to provide services over the internet. Cloud computing influences some technologies such as SOA to data security. In future, some of industry and vendors are expecting changes in IT trends and processes. We discuss here the data security & future of cloud computing. We will also review of current services provided by cloud computing in different arena.

**Keywords** Cloud computing · SOA

## 1 Introduction

The cloud computing is rising so established technologies day by day. The Small and Medium Business (SMB) are continuously increasing use of cloud being storing their data. There are decreasing the cost of maintaining and purchasing the infrastructure being cloud computing [1]. As far as cloud computing is concern, there are a lot of IT executives and information officers are worried about IDC related security. That's why multiple customers want such providers who have share different network and hardware resources. Some of country have privacy rule and regulations that prohibit the transmission of personal data out of country [2]. There is guaranteed quality

---

D. Gahlot (✉) · S. Tejasvee · K. B. Ranga

MCA Department, Government Engineering College Bikaner, Bikaner, Rajasthan 334001, India  
e-mail: [dr.devendragahlot@gmail.com](mailto:dr.devendragahlot@gmail.com)

S. Tejasvee

e-mail: [drsanjaytejasvee@gmail.com](mailto:drsanjaytejasvee@gmail.com)

K. B. Ranga

e-mail: [kunalranga@gmail.com](mailto:kunalranga@gmail.com)

R. R. Vyas

CSE Department, Government Engineering College Bikaner, Bikaner, Rajasthan 334001, India  
e-mail: [radhakrishnavyas@gmail.com](mailto:radhakrishnavyas@gmail.com)

of service to customizable, reliable, and dynamic by this environment [3]. There is no need to complete understanding of the infrastructure because users have heap of virtual resources. The founder of Sun microsystems declared the advection of cloud computing that “The network of computer” [4]. L Lori M. Kauffman examine some legal concerns, associated regulatory and issues of data security that came out as a primary enterprise computing platform being cloud computing emerges [5]. There are different areas to be at risk like multitenancy, external & internal storage security. The external storage, multitenancy, integration, lack of control, and dependency of data on the exclusive internet are required to secure. The large scale and facts are that kind of resources that belonging to cloud service providers which is heterogenous and distributed nature that is completely virtualized. In the present form of cloud is not enough for data authorization, identity, and authentication being traditional security mechanisms [6]. According to Ransome JF, it is great concern such as moving critical applications and sensitive data to public cloud environments for all the corporation which are continuously moving out of their data center’s network under their control. It is necessary to make ensure the data privacy and security of the customer by cloud solution provider and it is compulsory to provide all the evidence being data security and privacy for customers. It should also be on paper as service level agreements [7]. In this paper, we will describe the level of data security and the future scope against limitation of cloud computing.

## 2 Review Literature of Data Security and Future Issues

As far as our research is concern, we are describing data security and future issues of cloud computing.

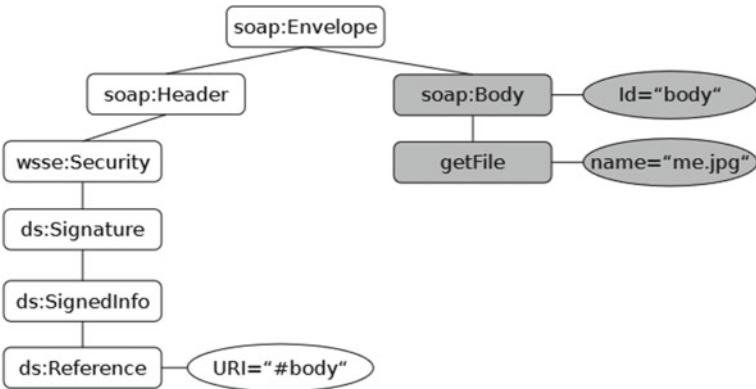
### 2.1 *Security Issues of Cloud Computing*

It is the major of data security for user. There are a lot of issues to described. Some of issues are described in this paper.

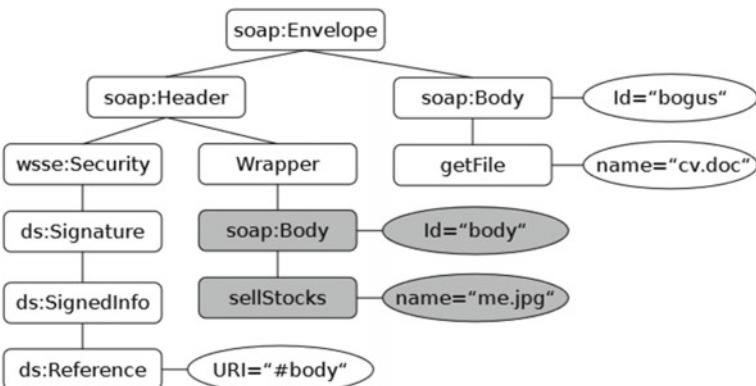
#### 2.1.1 **Signature of Extensible Markup Language**

The Extensible Markup Language signature element wrapping is used to prevent the attacks on protocols for integrity protection or authentication [8].

The wrapping of attacks is illustrated in the Figs. 1 and 2. The legitimate client sent SOAP message that is demonstrated in Fig. 1. The body of SOAP contains a request for the file “me.jpg” that is signed by the user. The value of the body attribute is having the signed message fragment using a Xpointer. The attacker may perform the attack such as, the actual body of request page may be moved into wrapping



**Fig. 1** Example SOAP message with signed SOAP body [9]

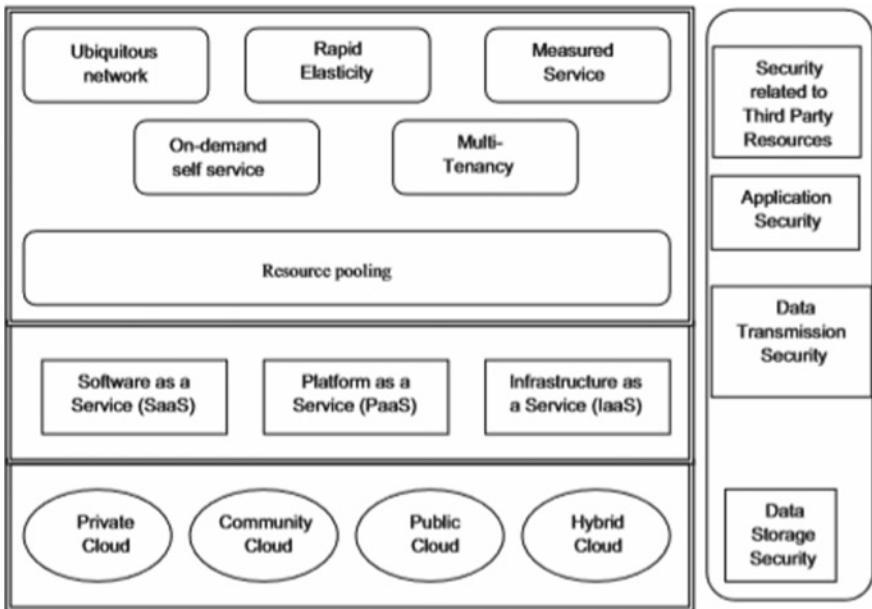


**Fig. 2** Example SOAP message after attack [9]

element which is newly inserted with creating new body. In Fig. 2 the file named “cv.doc” can have request from original sender’s authorization which is desire of attackers. The valid signature of a legitimate user will have resulting message [9].

The Schaad, Rahaman, and Rits said in their research paper titled “Towards secure SOAP message exchange in a SOA,” that the wrapping attacks by the attackers have variations and countermeasures being circumventing published [10]. According to L. Lo Iacono and N. Gruschka the real life wrapping attack had not been public till 2008 due to minimal uses of WSSecurity in business application. It was invented when the services of Amazon’s EC2 being their veneration to attacks’ wrapping [11].

S. Subashini and V. Kavitha have described in their research article titled “A survey on security issues in service delivery models of cloud computing” in Fig. 3. In this figure they said that the different deployment models of cloud namely hybrid, public, private, and community cloud are deployment models. There are different delivery models described above deployment layer. There are Infrastructure as a Service



**Fig. 3** Complexity of security in cloud environment [12]

(IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). All these models exhibit certain characteristics such as rapid elasticity, ubiquitous network, multitenancy and on demand self service which are shown in the top layer. There are a lot of security required by cloud which varies and depends with respect to the deployment model being used. In this research paper, it has also been described that Software as a Service model is used for deployment of software that is provided the licenses to customers for use as service on as per requirement. The Salesforce.com CRM application is the example of Software as a Service. Infrastructure as a Service is the delivery of infrastructure of computer as a service rather than purchasing network equipment, or data center space, software, and server. The client do outsourced services instead of buy [12].

### 2.1.2 Security of Internet Browser

The cloud is accessed from remote servers by the internet browser that is used only for input/output and also for authorization and authentication of commands to the cloud. It is observed that it has been categorized under different names such as SaaS, Web 2.0, and Web Application for last year's [9].

### 2.1.3 Data Binding and Integrity Issues of Cloud

Data integrity and security means that the data can be modified or accessed by authorized user to do so. According to Balachandra Reddy Kandukuri, Ramakrishna Paturi V, Dr. Atanu Rakshit it is process of verifying data. The guarantee of unmodified, correct, high quality data given by data integrity [13].

## 2.2 Future Issues

As far as Kim-Kwang Raymond Choo's research is concern the future of cloud may be seen as a pool of virtualized computing resources which may be permitted client to specific access to data and application in web-based environment on demand. They also described in their research paper that the cloud usage different model as well as different architecture. The public and private partnerships are developed as per need for cyber security's culture [14].

## 3 Conclusion

As far as the data security and future of cloud are concern cluster techniques may be more optimized then fog computing that may be called smog computing with new technique for data security including hosting malicious data, key cracking, password cracking, solving captcha as dynamic security code.

## References

1. Chen D, Zhao H (2012) Data security and privacy protection issues in cloud computing. In: IEEE, pp 647–651
2. Leavitt N (2009) Is cloud computing really ready for prime time? In: IEEE Computer Society, pp 15–20, January 2009
3. Wang L et al (2008) Scientific cloud computing: early definition and experience. In: Proc. 10th Int'l Conf. High-Performance Computing and Communications (HPCC 08). IEEE CS Press, pp 825–830
4. Urquhart J (2009) The biggest cloud-2. Computing issue of 2009 is trust. C-Net News, 7 Jan 2009. [http://news.cnet.com/8301-19413\\_3-10133487-240.html](http://news.cnet.com/8301-19413_3-10133487-240.html)
5. KaufMan LM (2009) BAE systems, “data Security in the World of Cloud Computing”. In: IEEE pp 61–64, July/August 2009
6. Li W, Ping L (2009) Trust model of enhance security and interoperability of cloud environment. Springer, Heidelberg, Beijing, pp 69–79
7. Ransome JF, Rittinghouse JW (2009) Security in cloud. In: Cloud computing. Implementation management and security. CRC Press
8. McIntosh M, Austel P (2005) XML signature element wrapping attacks and countermeasures. In: SWS '05: Proceedings of the 2005 workshop on Secure web services. ACM Press, pp 20–27

9. Jenson M, Schwenk J, Gruschka N, Iacono LL (2009) On technical security issues in cloud computing. In: IEEE 2009, pp 109–116
10. Rahaman MA, Schaad A, Rits M (2006) Towards secure SOAP message exchange in a SOA. In: SWS '06: Proceedings of the 3rd ACM workshop on Secure Web Services. ACM Press, pp 77–84
11. Gruschka N, Iacono LL (2009) Vulnerable cloud: SOAP message security validation revisited. In: ICWS '09: Proceedings of the IEEE international conference on web services. IEEE, Los Angeles, USA
12. Subashini S, Kavitha V (2010) A survey on security issues in service delivery models of cloud computing. Science Direct (Article), pp 1–11, July 2010
13. Kandukuri BR, Ramakrishna Paturi V, AtanuRakshit (2009) Cloud security issues. In: Proceedings IEEE international conference on services computing, September 2009
14. Raymond Choo K-K (2010) Cloud computing: challenges and future directions. Australia's National Research and Knowledge Centre on Crime and Justice, vol 400, Oct 2010

# Weather Event Prediction Using Combination of Data Mining Algorithms



**Yogesh Kumar Jakhar, Nidhi Mishra, and Rakesh Poonia**

**Abstract** Weather event prediction offerings suitable from the obsolete occurrences as a main gigantic obligation, since it depends on upon dissimilar constraints to forecast the destitute factors like air temperature, humidity, precipitation, wind speed, and dampness, which are fluctuating intermittently. A multi-model data mining approach is a unique process for merging the prognostic capability of multiple prototypes for better forecasting accuracy. In this paper, we proposed multi-model ensemble for forecasting weather events. The data mining algorithms Random forest, C5.0, AdaBoost, and Support Vector Machine (SVM) models are implemented in combination as ensemble. The combinations of (RF + SVM + AdaBoost) perform better accuracy with 82.73% in compare with other combinations of multi-model ensembles. For experimental work we used, weather data of Barajas Airport, Madrid, between 1997 and 2015 were gathered from web <https://www.wunderground.com/> The Weather Company, LLC.

**Keywords** Data mining · Multi-model ensemble · AdaBoost · Random forest

## 1 Introduction

Weather events are existence an accidental singularity its forecast has been always a dare for the meteorologist in the world everywhere. The various existing approaches for estimating this weather grounded on distinctive data collected by various means.

---

Y. K. Jakhar (✉)

Department of Computer Engineering, Poornima University, Jaipur, India  
e-mail: [yogeshjakhar@gmail.com](mailto:yogeshjakhar@gmail.com)

N. Mishra

Poornima University, Jaipur, India  
e-mail: [nidhi.mishra@poornima.edu.in](mailto:nidhi.mishra@poornima.edu.in)

R. Poonia

MCA Department, Government Engineering College, Bikaner, Rajasthan 334004, India  
e-mail: [rakesh.ecb98@gmail.com](mailto:rakesh.ecb98@gmail.com)

India is a country in which agriculture plays a predominant role both for employment and national income. Most of the agricultural products are needed for food for consumption and for industries as raw materials. Industrial development is needed for economic growth. Water is the most important resource for human being and for industries [1–4]. In today's information technology era, weather forecasting has become the most challenging and important technique which helps us to predict the atmosphere of a location. Weather estimation is an important application in meteorology and has become one of the furthermost systematically and technologically challenging problems for meteorologists around the world. From the last few decades, the advancement and development in science and technology enable scientists to make better and precise weather prediction for agricultural point of view [5–7]. Since ancient times, weather forecasting has been one of the most interesting and challenging area. One of the most important parameter of weather forecasting is rainfall estimation which is important for food production plan and water resource management. Rainfall plays an important role in agriculture and in other areas of society [8]. Among all data mining algorithm, the random forest is an adaptable machine learning algorithm and it executes or implements both regression and classification duties. C5.0 is a data mining algorithm, which was designed by Ross Quinlan, and the main use of C5.0 is to construct a decision tree. The AdaBoost model was used in grouping with different data mining algorithms to increase their forecasting capacities. A support vector machine (SVM) is a data mining model with accompanying knowledge algorithms, which is used to analyze data for classification as well as regression analysis. In SVM, first prearranged a set of training samples and then each marked as fitting to one or the other of two groups [9–13].

The structure of paper is systematized as follows. Associated work with data mining algorithms applications for weather events prediction especially rain prediction using single and multi-model combination as ensemble work is presented in Sect. 2. Section 3 consists with data details and preprocessing, and Sect. 4 has experiment evaluation process. Section 5 presents result analysis of the experiment.

## 2 Review of Literature

In this division, we briefly analyzed some research work related to weather event prediction. B. Narayanan and Dr. M. Govindarajan study the time series analysis for rainfall prediction using Support Vector Machine and Naive Bayes techniques. They proposed new techniques AdaSVM and AdaNaive using AdaBoost technique. Proposed AdaSVM and AdaNaive have given 98.66% and 97.62% of prediction accuracy, respectively [1]. Razeef Mohd, Muheet Ahmed Butt, and Majid Zaman Baba explored various data mining algorithms including decision tree-based J48, Random forest, Naive Bayes, Bayes Net, Logistic Regression, IBk, PART, and bagging and performed an experimental using weather data of Srinagar, India from November 2015 to November 2016. The Random Forest produces best rainfall prediction results with an accuracy of 87.76% [5].

Nazim Osman Bushara and Ajith Abraham examined relationship of rainfall with important parameters such as station, wind direction, date, humidity, minimum temperature, maximum temperature, and wind speed in Sudan. Authors have been used monthly meteorological data by Central Bureau of Statistics Sudan from 2000 to 2012 for 24 meteorological stations. The experiment shows that Date, Min-T, Humidity, and Wind D affect rainfall in Sudan [8]. Gabriela Grmanova and et al. proposed an incremental mixed collaborative model for time series forecast. The proposed system was able to forecast the power load. The achieved results show that the designed method could be a potential bearing in the superior of estimation models for time series with individual features [14]. Bohdan M. Pavlyshenko studies the application of machine learning algorithms for sales predictive analytics. Author founded that uses of regression for sales forecasting give better results with compared to time series methods [15]. Kanchan Bala, Dilip Kumar Choubey, and Sanchita Paul studied many research papers on the thunderstorm and lightning, which were based on soft computing and data mining methods. These techniques were neural network, fuzzy logic, rough set, genetic algorithm, SVM, k-means clustering, k-NN, etc. Author also suggested future research direction on these methods [16].

Vertika Shrivastava, Sanjeev Karmakar, and Sunita Soni studied a wide-ranging review of various assistances from 1997 to 2017. The BPN was appropriate to resolve this compound problem. The results showed 90% accuracy in modeling [17]. S. Karthick, D. Malathi, and C. Arun analyzed the performance of algorithms that were suitable for weather prediction. They compared prediction accuracy of C4.5 algorithm with Random Forest algorithm. The C4.5 succeeded an accuracy of 82.4%, the Random Forest was able to secure 87.1% accuracy proving it to be better [18]. Ali Ghasemy Armaki, Mir Feiz Fallah, Mahmoud Alborzi, and Amir Mohammadzadeh study the concept of hybrid models through machine learning algorithms. There were binary types of hybridization methods customary and collaborative methods. The proposed model was established on the customary mixture model of classification and clustering in which the stacking collaborative method was engaged in the classification part. In this paper, a novel framework was proposed for hybrid meta-learning to improve the predictive performance of credit scoring models. Based on the selected datasets, the results show that the hybrid meta-learner model of (KNN-NN-SVMPSO)-(DL)-(DBSCAN) outpaces all the literature's baseline classifiers in terms of accuracy rate and type I/II errors. This model also outperforms the best models used in the relevant literature in terms of accuracy rate with a significant margin [10]. Priyanka H U and Vivek R proposed multi-model prescient architecture, which was a unique methodology for joining the prescient capacity of different models for better forecasting. Results demonstrate that the proposed multi-model prescient architecture had the option to give preferable exactness over best model methodology. By displaying the blunder of prescient models, we had the option to pick subset of prototypes which yields precise outcomes. More data were displayed into framework by staggered mining which has brought about upgraded prescient exactness [11].

Changhyun Choi and et al were created forecasting models of hefty rain destruction using machine learning techniques for the city Seoul, and it is the Capital area of

the Republic of Korea. They utilized data on the event of substantial downpour harm from 1994 to 2015 as reliant factors and climate huge information as logical factors. The created model was established by applying AI systems, for example, decision trees, bagging, boosting, and random forests. Because of assessing the forecast exhibition of each model, the AUC estimation of the boosting prototypical utilizing meteorological information from the previous 1 to 4 days was the most astounding at 95.87% and was chosen as the final model [12]. Weiwei Lin, Ziming Wu, Longxin Lin, and Angzhan Wen And Jin Li (2017) proposed a collaborative random forest algorithm that routines the equivalent processing ability and memory cache system improved by Spark. The trial result demonstrates that the ensemble random forest algorithm beat SVM and other characterization algorithm in both execution and precision inside the imbalanced dataset [13].

### 3 Data Detail and Preprocessing

For experimental work, weather data of Barajas Airport, Madrid, between 1997 and 2015 were gathered from web <https://www.wunderground.com/> The Weather Company, LLC. The data in the dataset are in.CSV file. The dataset has 22 attributes and 6812 records. The following table, Table 1 showing the summary of numerical parameters of the dataset.

To deal with missing values, we used Amelia package in R programming to deal with missing values. Amelia package implements multiple imputations to deal with missing values. Multiple imputations help to reduce bias and increase efficiency. The outliers were replaced with maximum frequency value of parameter. Some parameters have categorical values, need to convert them into numerical values. The categorical values are replaced by numeric value for experiment purpose.

### 4 Experimental Details

The experiment work is carried out by dividing the datasets into two parts: training dataset and testing dataset. We used 1997–2014 years data for training purpose and 2015 data for testing purpose. Random forest, C5.0, AdaBoost, and Support Vector Machine (SVM) models are implemented in combination as ensemble. Experiments are carried out using R programming and also used R package for implementing these algorithms in combinations as ensemble. 16 types of different weather events are considered in the experiment according to dataset and tried to predict one on them, these are as follows: Normal day, fog, fog-rain, fog-rain-snow, fog-snow, fog-rain-thunderstorm, fog-thunderstorm, Rain, Rain-hail, rain-hail-thunderstorm, rain-snow, rain-thunderstorm, rain-snow-thunderstorm, thunderstorm, snow, and tornado. The following confusion matrix in Table 2, is constructed using normal day versus weather events, and events may be any one of above mentioned weather events.

For year 2015, 365 records are predicted as testing dataset.

**Table 1** Summary of numerical parameters of dataset

S. No.	Name of parameter	Minimum value	Maximum value	Mean value	Standard deviation
1	Max. Temperature	0	41	21.04	8.867
2	Mean. Temperature	-3	32	14.66	7.581
3	Min. Temperature	-10	28	8.639	6.838
4	Dew. Point	-12	20	8.118	4.743
5	MeanDew. Point	-15	16	4.974	4.655
6	Min. Dewpoint	-22	14	1.45	4.911
7	Max. Humidity	16	100	81.14	17.53
8	Mean. Humidity	15	100	57.97	19.67
9	Min. Humidity	4	100	34.73	19.32
10	Max. Sea Level Pressure	994	1047	1021	6.236
11	Mean Sea Level Pressure	986	1043	1018	6.48
12	Min. Sea Level Pressure	965	1041	1015	6.945
13	Max. Visibility	-10.6	47.62	14.52	8.58
14	Mean Visibility	-1.394	34.39	11.89	5.484
15	Min. Visibility	-1.497	31	9.479	5.034
16	Max. Wind Speed	0	182	21.95	9.904
17	Mean Wind Speed	0	39	9.171	5.11
18	Max. Wind Speed	3.588	156	39.97	12.37
19	Precipitation	0	32	0.1112	0.9672
20	CloudCover	-2.452	8	2.972	1.812
21	WindDirDegrees	-1	360	197.2	119.9

**Table 2** Confusion Matrix and Statistics for various combination of multi-model ensemble

Ensemble model	Weather event predicted	Normal day	Other than normal day
RF + SVM + C5.0	Normal day	245	40
	Other than normal day	35	45
RF + SVM + AdaBoost	Normal day	253	36
	Other than normal day	27	49
C5.0 + SVM + AdaBoost	Normal day	246	43
	Other than normal day	37	39
RF + AdaBoost +C5.0	Normal day	251	41
	Other than normal day	29	44

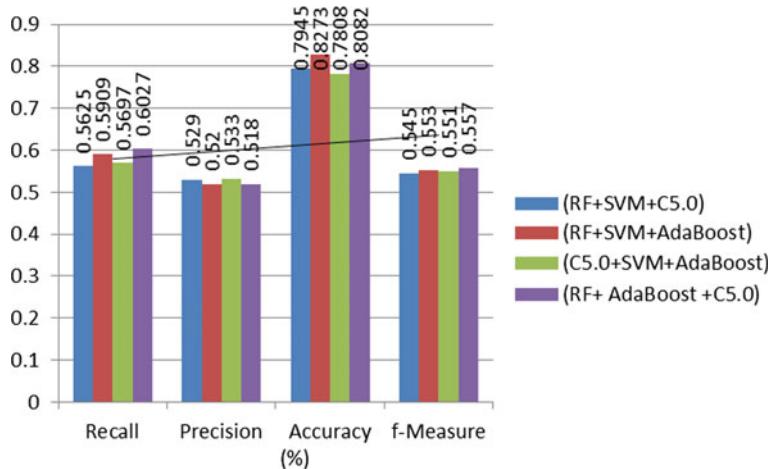
## 5 Results Analysis

Random forest, C5.0, AdaBoost, and Support Vector Machine (SVM) models are implemented in combination as ensemble. The combinations are (RF + SVM + C5.0), (RF + SVM + AdaBoost), (C5.0 + SVM + AdaBoost), and (RF + AdaBoost + C5.0). The following Table 3 has accuracy, recall, precision and f-measure of various combined models.

The combinations of (RF + SVM + AdaBoost) perform better accuracy with 82.73% in compare with other combinations. (RF + SVM + C5.0) perform 79.45% accuracy, (C5.0 + SVM + AdaBoost) perform 78.08% accuracy, and (RF + AdaBoost + C5.0) perform 80.82% accuracy. The following figure, Fig. 1 showing different performance measures of different combinations of data mining algorithms.

**Table 3** Performance measures of algorithms

Ensemble Models	Recall	Precision	Accuracy (%)	f-Measure
(RF + SVM + C5.0)	0.5625	0.529	79.45	0.545
(RF + SVM + AdaBoost)	0.5909	0.520	82.73	0.553
(C5.0 + SVM + AdaBoost)	0.5697	0.533	78.08	0.551
(RF + AdaBoost + C5.0)	0.6027	0.518	80.82	0.557



**Fig. 1** Different performance measures

## 6 Conclusion

Weather events prediction offerings suitable from the obsolete occurrences as a main gigantic obligation, since it depends on upon dissimilar constraints to forecast the destitute factors like air temperature, humidity, precipitation, wind speed, and dampness, which are fluctuating intermittently. A multi-model data mining approach is a unique process for merging the prognostic capability of multiple prototypes for better forecasting accuracy. For experimental work, weather data of Barajas Airport, Madrid, between 1997 and 2015 were gathered from web <https://www.wunderground.com/> The Weather Company, LLC. The experiment work is carried out by dividing the datasets into two parts: training dataset and testing dataset. The dataset has 22 attributes and 6812 records. Random forest, C5.0, AdaBoost, and Support Vector Machine (SVM) models are implemented in combination as ensemble. The combinations are (RF + SVM + C5.0), (RF + SVM + AdaBoost), (C5.0 + SVM + AdaBoost), and (RF + AdaBoost + C5.0). The combinations of (RF + SVM + AdaBoost) perform better accuracy with 82.73% in compare with other combinations. (RF + SVM + C5.0) perform 79.45% accuracy, (C5.0 + SVM + AdaBoost) perform 78.08% accuracy, and (RF + AdaBoost + C5.0) perform 80.82% accuracy.

## References

1. Narayanan B, Govindarajan M (2016) Rainfall prediction based on ensemble model, vol 5, issue 5, pp 8237–8243, May 2016
2. Moreale P, Holtz S, Goncalves A (2013) Data mining and analysis of large scale time series network data. In: 27th international conference on advanced information networking and applications workshops, IEEE, 978-0-7695-4952-1/13
3. Dhore A, Byakude A, Sonar B, Waste M (2017) Weather prediction using the data mining Techniques. Int Res J Eng Technol (IRJET) 04(05):2562–2565
4. Piruthvei C, Kanimozi Selvi CS (2017) Filtering of anomalous weather events over the region of Tamil Nadu. In: IEEE international conference on intelligent techniques in control, optimization and signal processing, 23–25 Mar 2017, Srivilliputtur, INDIA, 978-1-5090-4778-9/17
5. Mohd R, Butt MA, Baba MZ (2018) Comparative study of rainfall prediction modeling techniques (A case study on Srinagar, J&K, India). Asian J Comput Sci Technol 7(3): 13–19
6. Mazhar A, Ikram MT, Butt NA, Butt YJ (2015) Do we really have to consider data mining techniques for meteorological data. In: Fourth International Conference on Aerospace Science and Engineering (ICASE), IEEE. <https://doi.org/10.1109/icase.2015.7489525>
7. Nagalakshmi R, Usha M (2013) Application of data mining techniques in maximum temperature forecasting: a comprehensive literature review. Int J Adv Res Comput Sci Manage Stud, Special Issue, pp 1–9, December 2013
8. Bushara NO, Abraham A (2014) Weather forecasting in Sudan using machine learning schemes. J Netw Innov Comput 2:309–317
9. Navaz S, Khan H, Ghosh SM (2017) A survey on ensemble computing method for rainfall prediction in different regions of Chhattisgarh. Int J Sci Res (IJSR) 6(6):19–25
10. Armaki AG, Fallah MF, Alborzi M, Mohammadzadeh A (2017) A hybrid meta-learner technique for credit scoring of banks' customers. Eng Technol Appl Sci Res 7(5):2073–2082
11. Priyanka HU, Vivek R (2016) Multi model data mining approach for heart failure prediction. Int J Data Mining Knowl Manage Process (IJDKP) 6(5):31–37

12. Choi C, Bae Y, Kim J, Kim HS (2018) Development of heavy rain damage prediction model using machine learning based on big data, Hindawi. In: Advances in meteorology, vol 2018, Article ID 5024930, 11p
13. Lin W, Wu Z, Lin L, Wen A, Li J (2017) An ensemble random forest algorithm for insurance big data analysis. IEEE Access, Special Section On Recent Advances In Computational Intelligence Paradigms For Security And Privacy For Fog And Mobile Edge Computing 5:16568–16575
14. Grmanová G, Laurinec P, Rozinajová V, Ezzeddine AB, Lucká M, Lacko P, Vrablecová P, Návrat P (2016) Incremental ensemble learning for electricity load forecasting. Acta Polytechnica Hungarica 13(2):97–117
15. Pavlyshenko BM (2019) Machine-learning models for sales time series forecasting, MDPI. Data 4:15. <https://doi.org/10.3390/data4010015>
16. Bala K, Choubey DK, Paul S (2017) Soft computing and data mining techniques for thunderstorms and lightning prediction: a survey. In: International Conference on Electronics, Communication and Aerospace Technology (ICECA 2017), IEEE, 978-1-5090-5686-6/17
17. Shrivastava V, Karmakar S, Soni S (2017) Suitability of neural network for weather forecasting: a comprehensive literature review. Int J Recent Sci Res 8(12):22300–22316
18. Karthick S, Malathi D, Arun C (2018) Weather prediction analysis using random forest algorithm. Int J Pure Appl Math 118(20):255–262

# Data Compression and Visualization Using PCA and T-SNE



Jyoti Pareek and Joel Jacob

**Abstract** This paper examines two commonly used data dimensionality reduction techniques, namely, PCA and T-SNE. PCA was founded in 1933 and T-SNE in 2008, both are fundamentally different techniques. PCA focuses heavily on linear algebra while T-SNE is a probabilistic technique. The goal is to apply these algorithms on MNIST dataset and to see how they practically work and what conclusions we could draw from their application. The objective is to reduce the dimension of the data while retaining most of the information. We perform both these techniques and make a comparison between them by observing the results. We note the behavior of the reduced components obtained from both techniques, by visualizing it in 2-dimensional space. Upon further research and application, it became apparent that the data dimensionality reduction is sensitive to the parameter settings and must be fine-tuned carefully to be successful.

**Keywords** Dimensionality reduction · Visualization · Reduced components

## 1 Introduction

Nowadays, datasets have large number of features along which data are distributed; in other words, we have large number of dimensions to be explored in order to analyze the data. The problem is some of the features are redundant and do not give much insightful information about the dataset. Redundant features not only make the analysis of data difficult but also increase the processing time, along with consumption of memory which could have been utilized otherwise. Also, visualization of higher dimension data is a problem since it becomes difficult to visualize beyond 3 dimension. Visual exploration of data is really important to understand the nature of the data. Therefore, our solution is to first compress the data and then visualize it.

---

J. Pareek (✉) · J. Jacob  
Department of Computer Science, Gujarat University, Ahmedabad, India  
e-mail: [drjyotipareek@yahoo.com](mailto:drjyotipareek@yahoo.com)

Compression and visualization of data can be achieved using dimensionality reduction techniques. Here, we will focus on two such techniques, namely, PCA and T-SNE.

Principal component analysis is a statistical technique that is useful for compression and visualization of data. It is one of the oldest and most used dimensionality reduction technique. The goal of this technique is to find the direction of maximum variance, and this direction is given by eigen vectors of the co-variance matrix, which is calculated from the given dataset. These eigen vectors are also called the principal components. The data points are then projected on the principal components, which explain the most variance, thereby reducing the dimension. The dimension of the dataset is reduced in such a way that the reduced dataset retains most of the information. T-Distributed Stochastic Neighbor Embedding is a relatively new technique compared to PCA. It is a probabilistic technique and is well suited for visualization of higher dimensional data. The goal of T-SNE is to find an optimal way to project data points into a lower dimensional space such that closeness or clustering of these data points is similar as were in high-dimensional space. In essence, it tries to maintain the internal structure of the data in low-dimensional space.

## 2 Literature Review

Dimensionality reduction is an essential pre-processing step in many machine learning applications. The main goal of the dimensionality reduction algorithms is to transform the data or features from a higher dimensional space to a lower dimensional space [1]. Dimensionality reduction techniques improve the overall performance of the algorithms [2]. Among various dimensionality reduction techniques, we decide to explore PCA and T-SNE in this paper.

PCA is a popular dimensionality reduction algorithm that is based on projecting or mapping a higher dimensional data points to lower dimensional space [3]. The principal components are those vectors that explain the amount of information or variance of the dataset in question. Hence, the number of principal components to be considered depends upon the user's requirement [4]. PCA not only explains the behavior of each independent variable but also provides information of how much the input variables are correlated with each other [5]. Usually for plotting purpose, two or three principal components are enough, as beyond that the visualization becomes difficult to interpret, but in case of building a model, the number of principal components should be carefully chosen in terms of their significance [6].

T-SNE is relatively a new algorithm compared to PCA for the visualization of data that is capable of maintaining the global structure of the data and at the same time keeping the local structure intact [7]. This algorithm provides a slight variation over the Stochastic Neighbor Embedding [8].

One of the challenges faced for the accurate visualization of high-dimensional data is the dissimilarity of shapes between clusters of high dimension and clusters formed in low dimension. This challenge was tackled using t-SNE, by taking into

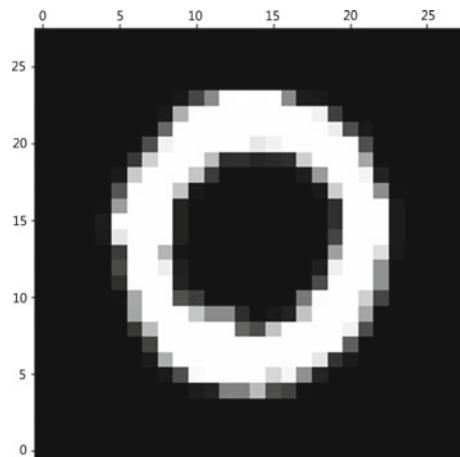
consideration the local and global structures of the large dataset at different scales. The global structure consists of geometrical features, and local structure is the pixel-based information of an image [9]. Initially, the algorithm places all the points in 2-dimensional plane at random positions and allows them to interact as if they are two physical particles. The interaction is governed by two laws: first, all points are repelled from each other; second, each point is attracted to its nearest neighbor. The most important parameter of t-SNE is perplexity, which defines how many points each data point should consider as its neighbor and be attracted to so that structure of points is intact even in lower dimensional space [10]. The graph of t-SNE will change during every run as t-SNE transforms the samples into different spaces that preserve distances between them and not the value of the data sample [11].

### 3 MNIST Dataset

The MNIST dataset consists images of handwritten digits. This dataset consists of 70,000 images, which are divided into train set and test set. The train set consists of 60,000 images, which are used to train the model, while the test set has remaining 10,000 images.

Although we have just considered 20,000 examples, as our primary aim is to just compress and visualize the data in lower dimensions. It is an image of  $28 \times 28$  pixels, which equals to 784 dimensions. The dataset consists of such handwritten digits from 0 to 9 in various handwritings.

Here is a sample from the dataset (Fig. 1).



**Fig. 1** A sample image from MNIST dataset

## 4 Methodology

Let's see what are the steps to perform each algorithm. We will first start with PCA.

### 4.1 Principal Component Analysis

- Consider a dataset of  $M \times N$  dimension.
- Compute the mean and subtract it from its corresponding data value for every dimension of the whole dataset. The resultant dataset is denoted as  $B$ .
- Calculate the co-variance matrix of the whole dataset.

$$S = \frac{1}{N} B^T B \quad (1)$$

- Calculate the eigen values and its corresponding eigen vectors of the co-variance matrix computed in the previous step.
- Sort the eigen vectors in decreasing eigen values and choose  $z$  eigen vectors with the largest eigen values to form a  $d \times z$  dimensional matrix  $W$ .
- Use this  $d \times z$  eigen vector matrix to project the samples onto the new subspace.

$$Y = x^T W \quad (2)$$

The original dataset, before applying PCA, with 784 dimension. For convenience, we will only be showing first 10 rows of the dataset here (Fig. 2).

After applying PCA, 784 principal components are obtained since there are 784 dimensions hence 784 eigen values. Out of 784, we choose  $k$  components that carry maximum information by sorting them in decreasing eigen values. Since we are interested to visualize in 2 dimension, we will only choose top 2 principal components,

	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel775	pixel776	pixel777	pixel778	pixel779	pixel780	pixel781	pixel782	pixel783
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

**Fig. 2** The original dataset with 784 dimensions

	1st_principal	2nd_principal	label
0	-5.558661	-5.043558	1.0
1	6.193635	19.305278	0.0
2	-1.909878	-7.678775	1.0
3	5.525748	-0.464845	4.0
4	6.366527	26.644289	0.0
5	-0.557059	1.201279	0.0
6	6.440129	-6.118906	7.0
7	4.421476	0.215520	3.0
8	-1.315634	-0.724664	5.0
9	-2.603522	3.106035	3.0

**Fig. 3** Reduced dataset with top 2 principal components

which explain the maximum variance. The reduced components are orthogonal and uncorrelated with each other (Fig. 3).

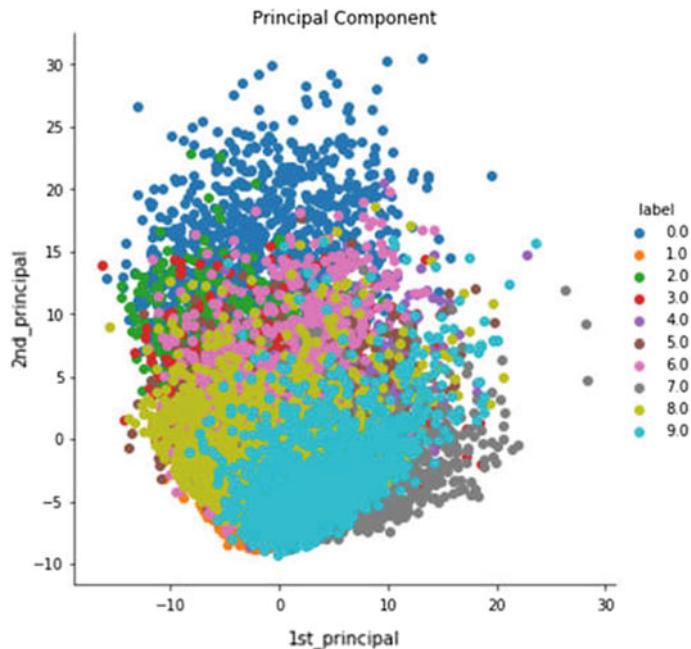
By reducing the original size of dataset into two dimension, we achieved the goal of data compression. Next, we visualize the reduced dimension data as to see how the data points behave in lower dimension (Fig. 4).

This is the resultant plot by just using the top two principal components. Clearly, the points are not well separated and are overlapping with each other. This could indicate that the chosen principal components might not be sufficient to explain the majority of variance of the data and we might need to consider the contribution of other principal components as well.

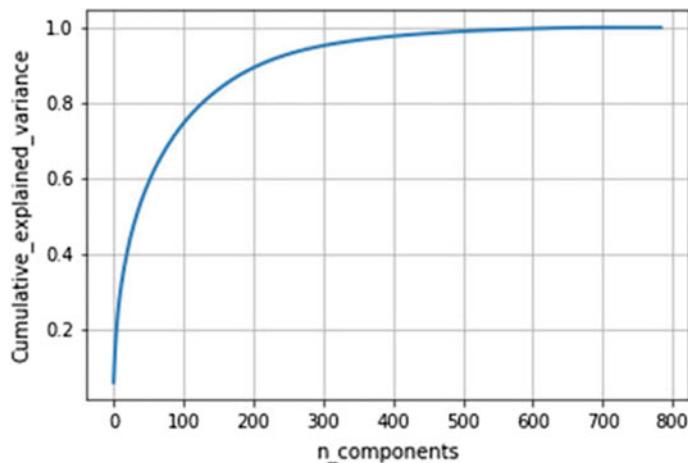
Now let's plot the graph between cumulative explained variance of principal components and total number of principal components as to get an idea of how many principal components should be considered to explain most of the variance (Fig. 5).

From the graph, it is evident that if we take 200 dimensions, approx. 90% of variance can be explained.

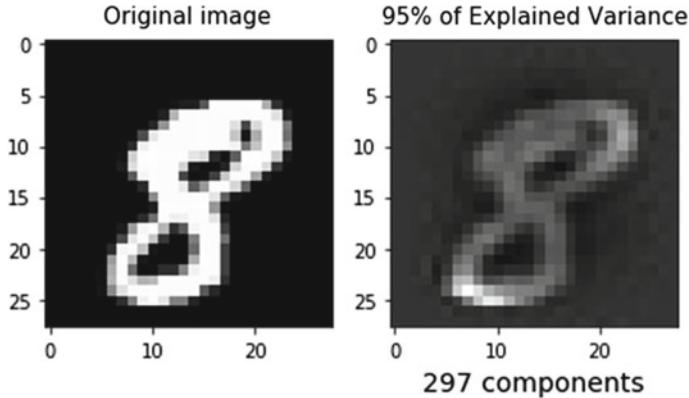
We use 297 components, which approximately explain 95% of variance. From the above figure, it can be seen that even though the compressed image is not as perfect as the original image, it still manages to explain the original image at its best (Fig. 6).



**Fig. 4** Visualization of dataset that was reduced using PCA



**Fig. 5** Cumulative explained variance versus number of principal components



**Fig. 6** Image with 784 dimensions versus image with 297 dimensions

## 4.2 T-Distributed Stochastic Neighbor Embedding

- Given a dataset of high dimension, t-SNE first measures pairwise similarities between two data points as follows:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}, \quad (3)$$

- As Van der Maaten and Hinton explained: “The similarity of data point to data point is the conditional probability, that would pick as its neighbors if neighbors were picked in proportion to their probability density under a Gaussian centered at.”

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad (4)$$

- Calculate the probability of points in the corresponding low-dimensional space.

$$q_{ij} \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}} \quad (5)$$

- Our aim is to make and identical so that structure in low-dimensional space will be identical to high-dimensional space. We use *KL* divergence that is a natural distance measure between probability distribution. We use gradient descent to move points around such that *KL* divergence is minimum. The cost function is given below:

	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel775	pixel776	pixel777	pixel778	pixel779	pixel780	pixel781	pixel782	pixel783
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

**Fig. 7** The original dataset with 784 dimensions

	Dim_1	Dim_2	label
0	72.716675	-26.153477	1.0
1	-48.665245	-43.919968	0.0
2	54.325542	22.210487	1.0
3	-40.087101	42.154762	4.0
4	-55.703262	-50.109032	0.0
5	-28.121956	-37.649361	0.0
6	4.174982	68.099297	7.0
7	7.242168	2.418088	3.0
8	20.208687	-38.560211	5.0
9	-9.152531	-16.008718	3.0

**Fig. 8** Reduced dataset using T-SNE

$$C = KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (6)$$

## Experimental Setup

Technique	Parameter
T-SNE	n_components = 2 Perplexity = 30 Learning rate = 200 Number of iterations = 1000

Before applying T-SNE, the original dataset with 784 dimensions (Fig. 7).

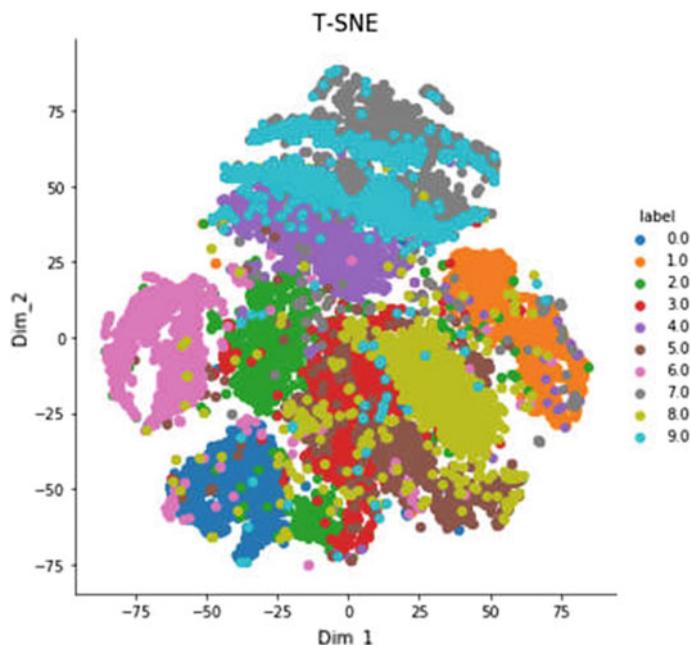
T-SNE builds a probability distribution over the points in high-dimensional space such that similar points will have a high probability of being picked with an aim to embed them in lower dimensional space, this probability is denoted as  $p_{ij}$ . On the other hand, dissimilar points will have an extremely low probability of being picked.

After that, a similar probability distribution is constructed over the points in the low-dimensional space, this probability is denoted as  $q_{ij}$ .

Once we get  $p_{ij}$  and  $q_{ij}$ , our aim is to make them identical. This is done by using gradient descent to minimize the  $KL$  divergence between the two distributions.

After performing T-SNE, the dataset has reduced to two dimensions as shown below (Fig. 8).

We visualize the reduced dimension calculated using T-SNE and observe how the data points behave when projected into low-dimensional space. We can obtain different configuration of plot by changing the hyperparameters like perplexity, learning rate, and number of iteration. The figure below we obtained is with optimum value of hyperparameter (Fig. 9).



**Fig. 9** Visualization using T-SNE

From the above figure, it can be seen that even though some of the points are overlapping, the visualization is much better as compared to PCA.

## 5 Observation

When the dimension of MNIST dataset was reduced to 2-D using PCA, the resultant clusters were not well separated and there were lot of overlapping between clusters, and hence, it was difficult to distinguish between different labels. When the cumulative variance ratio was calculated, it was observed that the top two principal components only contributed 10.15% of the total information, i.e., 5.91% explained by first principal component and 4.25% by second principal component.

For further investigations, we plotted the graph of ‘cumulative explained variance’ versus ‘number of components’ and it was inferred that to explain 90% of total information, at least 200 components should be taken into account. In the paper, 297 components are considered, which explain 95% of the variance.

On the other hand, when T-SNE was used to reduce the same dataset into 2D, it gave better results. Although there was some overlapping between the points, but the clusters were well separated compared to PCA. This is because unlike PCA, T-SNE tries to conserve the internal structure of the data, i.e., it tries to projects points on low-dimensional space while conserving the clustering structure of high-dimensional space.

## 6 Conclusion

PCA is a linear algorithm with an aim to maximize the variance and preserve the global structure of the data. That is, if there are two separate clusters in high-dimensional space, when projected in low-dimensional space, they are placed near to each other. Hence, we lose the information that those were two separate clusters. It was for the same reason, the plot in Fig. 4 was very much cluttered, and it was difficult to distinguish between various clusters.

While T-SNE performs well when dealing with non-linear manifold structures since it aims to conserve the local structure of the data. That is, the clusters are conserved even in low-dimensional space as it was in high-dimensional space that was evident in the plot of Fig. 9.

Hence, T-SNE is well suited for visualization of high-dimensional dataset since it very well comprehends the complex polynomial relationship between features, which PCA fails to do so since it is a linear algorithm.

## 7 Future Scope

After comparing and using both the algorithms, it was evident that both algorithms come with a cost. Even though T-SNE performs well for the visualization of high-dimensional data, it is computationally very expensive. It takes lot of time to process the data, which are not convenient in the real-world scenario while dealing with large dataset; on the other hand, PCA is computationally fast but loses on visualization domain.

In future, we can combine both the techniques successively on the same dataset. First, we perform PCA on the dataset thereby reducing the dimension to a significant number and then try to visualize it using T-SNE, so that with already decreased dimension, we can visualize better with less computational time. We can also explore other dimensionality reduction techniques such as Sammon's mapping, Low variance Filter, Decision Trees, High Correlation Filter and compare it with the techniques explained in this paper.

## References

1. Tharwat A (2009) Principal component analysis—a tutorial. In: Iunderscience enterprises
2. Aráujo D, DóriaNeto A, Martins A, Melo J (2011) Comparative study on dimension reduction techniques for cluster analysis of microarray data. In: International joint conference on neural networks
3. Jolliffe IT (1986) Principal component analysis and factor analysis. Springer, Berlin
4. Lefter C, Bratu G et al (2006) Marketing, vol 1. Transilvania University of Brasov Publishing House, Brasov. In
5. Qi X, Luo R (2014) Sparse principal component analysis in Hilbert space. Scandinavian J Statistics. <https://doi.org/10.1111/sjos.12106>
6. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. In: Chemometrics and intelligent laboratory systems, vol 2, no 1–3, pp 37–52
7. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9
8. Hinton G, Roweis S (2002) Stochastic neighbor embedding. In: Advances in Neural Information Processing (NIPS)
9. Husnain M, Missem MMS, Mumtaz S, Muzzamil M, Luqman MM, Coustaty M, Ogier J-M (2019) Visualization of high-dimensional data by pairwise fusion matrices using t-SNE. In: Symmetry
10. Kobak I D, Berens P (2018) The art of using t-SNE for single-cell transcriptomics. In: bioRxiv
11. Nanaware T, Mahajan P, Chandak R, Deshpande P, Patil M (2018) Exploratory data analysis using dimension reduction. IOSR J Eng (IOSRJEN)

# Appscrumfall: APP Development Methodology Based on ScrumFall



Prerna Bisaa

**Abstract** Agile is a combination of iterative and changeable process which is adopted by IT organization. All software projects are specialized. It is necessary to choose a software model to build any software but a single software development model may not be suitable for all types of projects. Developers face difficulties with existing models for the duration of development. They tackle challenges by balancing the software development lifecycle according to their needs. Therefore, in this paper, we have introduced a software process methodology that features both the scrum and waterfall approaches and is named “appscrumfall.” Software development of mobile app using the model ““Scrum-Fall” (scrum blended with waterfall)” is practicing in the company to solve the deficiencies of the traditional model. We have analyzed the performance and availability to implement this process model. The result shows that this process model is effective for software projects. The aim of this model is to provide a formula for better implementation in large IT sectors.

**Keywords** Scrumfall · Agile · Spiral · XP · RAD · Appscrumfall

## 1 Introduction

Software process model also referred as software development life cycle (SDLC) the sequence of required phases for the entire lifetime of a software artifact. This model covers everything from the origin of a project by communicating with clients until the phase-out of the software product. The goal of following a process model is to split the software development activities into distinguishable, unambiguous phases with the purpose of improved research and management to achieve economic success [1]. Early 1960s and then from the late 1960s to present, many models have been proposed and used in the industry to develop software in an effective manner [2]. Some of the commonly used models are linear model, sequential development model, insistent, incremental model, spiral development, rapid application development, prototyping,

---

P. Bisaa (✉)

Tantia University, ShriGanganagar, India

e-mail: [prema.bissa@gmail.com](mailto:prema.bissa@gmail.com)

scrum, kanban, and dynamic system development method. All these models have their strength and weakness based on their nature, and these models can be divided broadly into two categories: traditional and predictive, which is also called plan-driven and agile processes. In the case of a plan-driven process, all the activities for the entire lifetime of the product are preplanned and progress is calculated based on the plan [2]. As a result, it is difficult to adopt any changes in the middle of development. A plan-driven model is suitable when the product and the team are large, the product is highly critical and hard to scale down, and the development environment is stable. In a plan-driven model, experienced personnel are required only at the beginning of the project and success is achieved through structure and order. Despite that, if the development environment is dynamic, that means changes in requirements are occurring frequently, then it is expensive [3]. On the other hand, it is easier to change plans as planning is incremental in agile methodologies. As a consequence, changes in user's requirements can be easily adopted and reflected in the software. An agile process is suitable when the product and team are small in size and the development environment is dynamic in nature [4]. This paper presented a hybrid software process model which have named as "appscrumfall." This process model includes the elements of plan-driven and agile processes which facilitate to cope up with the dynamic as well as stable nature at the different phases of SDLC. The main contributions of this paper are as follows:

- Presentation of a practical hybrid software process model combining elements of plan-driven and agile processes.
- Possible characteristics of a software project that is suitable to apply this model.
- Verification of the value of this model using real-world data.

## 2 Related Work

The software process model is a foundation of complete software projects and applicable to all types of software process. Process model is examined by software process assessments, which leads to software process improvement and identifies capability and risks of the projects. In the past decade, many software process models were proposed; in this section, we discuss on pros and cons regarding efficient process models. Here, we have mainly included the process models that are commonly used in software industries [5].

### 2.1 Plan-Driven Development

It is a more traditional and formal specific approach for developing systems. This model tries to plan and predict user's requirements that might be wanted in the end product. In plan-driven development, specific phases are followed in a sequential

manner. This is divided into three categories: sequential, incremental, and evolutionary. Some of the most widely used plan-driven development models are briefly discussed below.

### **2.1.1 Waterfall**

This is a oldest standard for software engineering; this is also known as “the classic life cycle”; it is a linear with chronological approach to software development that begins with customer requirements. This model was proposed by Winston Royce. This model works fine on big and weak teams. However, the model lacks flexibility as the adaption to requirement change may cost time and more money. It can be provided as a useful process model. Normally, this model is a forward manner process. Analysis, requirement, design code test maintenance and deployment are its phase. In this analysis, the requirement is determined in the work itself, if any situation. If the user has a problem in design or code phase, then the client has to start again.

### **2.1.2 Incremental**

It is a mixture of the classic life cycle model and perspective model applied in an insistent mode. In this model, all the work is done incrementally, different types of increments are used, and the work is completed in a phased manner.

### **2.1.3 Rapid Application Development (Rad Model)**

It is an incremental and quickly functionalized model that emphasizes shorter development cycles and “high-speed” transformation of the waterfall model. Development is achieved using a component-based production approach in a rapid manner; hence, it is called rapid application development. It is noted to be a fully functional system with a very short time span. Like all process models, Red Approach has drawbacks for large and scalable projects. This is not appropriate when the technical risk is high. It has different teamwork so team co-ordination is necessary. Reds need enough resources to make the right number of Red teams.

### **2.1.4 Perspective Process Model**

These models are insistent and enable developers to develop more and complete versions of the software.

### 2.1.5 Prototyping Model

It is an evolutionary process model, and it is commonly used as a method that can be executed in the context of one of the process models. This helps customers and developers to understand better, and for this, some dummy projects are created in this model to introduce customers to what to create when requirements are complicated. For some time insufficient requirements analysis, confusion between customers between prototypes and prototypes between real systems can be problematic.

### 2.1.6 Spiral Model

There is a progressive process model proposed by Boehm [BOE88] which works like spiral. It carries validation and validation processes. The model has a prototypical nature with a controlled and systematic aspect of waterfall. It also focuses on risk management that is used to direct multi-stakeholders of a software intensive system. It has two main features: One is a cyclical approach to the incremental growth of the system, and the other is to ensure commitment with stakeholders along with satisfactory system solutions.

## 3 Agile Model

Agile is an approach to software delivery, which builds software from the start of a time-boxed, iterative project, rather than attempting to deliver it all together until the end. This model has many variants so briefly discussed below.

### 3.1 Scrum Model

It is a popular methodology of agile approach in which the software development is partitioned into three phases: One is planning and designing, another is sprints according to planning, and last is testing deployments.

### 3.2 Extreme Programming (XP)

The extreme programming model is originated from the disadvantages of the traditional process model. This process model addresses the primary risks in software development, and it is formed based on ordinary logic principles and understandable practices.

### ***3.3 Dynamic Systems Development Method***

DSDM has been developed from the expertise gained by a large group of vendor and user organizations. The main focus of DSDM is to deliver urgent business needs on time. It focuses on regular product delivery. It also facilitates the reversibility of changes at any time. DSDM has the potential risks of lack of user involvement or too much user involvement, extreme time on decision-making, and development of irreparable implements. The system may fail as testing is not integrated throughout the life cycle.

### ***3.4 Feature-Driven Development***

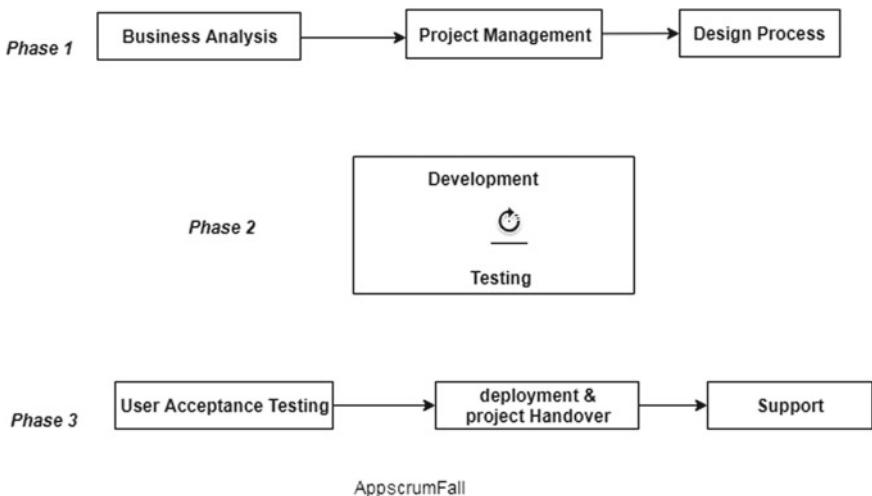
It is an agile software development process which uses iteration model [17]. In this model, an important feature is “feature.” This model was created to level down the larger projects and teams easily by combining of other well-known agile approaches and industry-recognized practices. The development process of FDD is divided into five process develop a model; build a features list, plan by feature, design by feature and build by feature. This model is mainly suitable for large systems and pays less attention to the initial design.

### ***3.5 Water-Scrum-Fall Model***

Water-Scrum-Fall model is a hybrid method where the software development activities are divided into three separate phases: Water, Scrum, and Fall. Water represents the upfront activities such as defining requirements and planning. Scrum is the middle phase of the process where the software is actually built. On the other hand, Fall controls the release of software release frequency by forming gates. Although the Water-Scrum-Fall model suggests limiting the time spent on upfront activities; however, there is no direction regarding how to deal with the mixed nature of environment during development.

## **4 Proposed Model**

The proposed appscrumfall model is a subset of scrum and waterfall model so it is also called hybrid manner model and combined for best app development practices. This model emphasizes the agility in a large projects and how to cope up with dynamic as well as stable development environment. This app development methodology is based on water scrum fall in which three variants business analysis project management



**Fig. 1** Appscrumfall flowchart

and design process is in first phase work as plan-driven manner. It addresses all requirements and designing, and it is a change-oriented phase in which requirements are accommodate. Next phase is scrum-based in which development and testing are done, and last phase is again incremental and insistent manner, where user acceptance testing deployment and feedback are done. Figure 1 demonstrates the different steps and activities of the proposed model.

## 4.1 Phase 1

### 4.1.1 Business Analysis

After the initial communication and agreement with the client, the actual work of developing a software starts with analyzing business, so this subpart b of first phase is to understand business problem and solution, and it follows objectives and guidelines of customer and stakeholders for best practices put efforts for maximum input and output for better results and in with integrated the requirements.

### 4.1.2 Project Management

Project management makes clarification and requirement clear for the undergone projects; it schedules project with given time frame and identify and quantify software quality attributes and their measuring scale or unit. Sign up the projects on the basis of requirements and time scheduling. Training on tools, technology or practices

if required. Product backlog creation with requirements at hand. Cost and effort estimation for each task in the product backlog list.

#### **4.1.3 Design Process**

At this subphase, the initial architecture of the system and high-level design for implementing the requirements from the product backlog list are constructed. Besides, identification of changes and refinement of architecture to implement new requirements, domain analysis, and risk analysis are also included in this subphase.

- set wire frames and story boarding
- Get insight into apps functional architecture
- Prototype design and testing
- Finally, converted to html/css.

### **4.2 Phase 2**

The main difference between the scrum model and proposed model is the development and testing phase. In this proposed model, development work is done in a mix of insistent and sequential manner. The development team down with the other stakeholders determines whether the development time and, at the end of iterations, quality or functionalities of the product are met as the requirements for the product delivery. As like the scrum model, this proposed model also consists of following macro processes:

Review release plans in the team meeting.

- Conformant of the sharing, analysis, and change of the standards.
- Several sprints until software delivery.

Sprint consists of development actions which are performed in both the insistent and sequential manner in the ScrumFall model. The earlier activities like communication and design, requirement analysis, specification and design, and prototyping are performed in an insistent manner in the Sprint, and this iteration continues until the client comes to an agreement with the developer. In this phase, we approach projects as sprints in which each has entry and exit points, and development and testing are carried out simultaneously with each new build of the app.

- Builded apps are tested for platform-specific, native, cross platform, and hybrid
- Emphasized actual binding of the app and its features
- Development and testing phase is more intensive.

### **4.3 Phase 3**

#### **4.3.1 User Acceptance Testing**

UAT is done by the client along with the testers so thoroughly evaluates the usability, functionality, and design of the apps, and clients are happy with this and certify the projects to the requirements and detect and remove existing and potential errors in this phase.

#### **4.3.2 Deployment and Project Handover**

In this phase, work is transferred to server on the clients approval and given 6 months postdelivery guarantee. These IP of app projects are 100% clients owned. Codes are completed, and documents are handed over or finally apps “GO Live”.

#### **4.3.3 Support**

Customized maintenance plans and provide customer support for 24/7.

#### **4.3.4 Objectives**

Although the presented model in this paper has similarities with scrum model; however, it can be distinguishable easily for the clear and comprehensive combination of the elements from plan-driven and agile processes. The major dissimilarity between the proposed model and existing scrum model is the nature of the sprint. As a result, facts that are observed by the development team in anonymous company after choosing ScrumFall model as their SDLC are as follows: The increment of each sprint such as application programming interface (API), framework, library, and software development kit (SDK) is used by another team for development. Some objectives are captured:

- Improvement in delivery time and less delay in future delivery.
- Increase in the number of sprints to be delivered.
- The increase in the number of fixing bugs.
- Accommodate changes in requirements without risking a delay in delivery.
- User involvement change when it is needed.
- Each feature is large and highly complicated.
- Geographically distributed large teams.
- The development environment is dynamic at earlier stages and then gets stable in later stages.
- The team combines both experienced and inexperienced personnel.
- User involvement throughout the SDLC is not possible or necessary.

## 5 Conclusion

As the software technology is elevating every moment, to achieve a successful project with efficiency, the software models need to be improved. On the specific types and scales of projects, previous plan-driven models and agile models have their success. For instance, these models are suitable for large and steady systems and requires only experienced personnel at the beginning of the project, and agile is more suitable for small systems and requires specialist agile personnel all over the project. From the practitioners' experience, appscrumfall holds the success over large, critical systems, geographically distributed large teams where the team is combined by both experienced and inexperienced personnel. In addition, appscrumfall has shown effectiveness in time, cost, and economic factors. In future, we plan to deploy this model on a large scale at dissimilar organizations to verify the further success of the proposed model.

## References

1. Alexandros NK, Sakas DP, Vlachos DS, Dimitrios NK (2017) Comparing scrum and XP agile methodologies using dynamic simulation modeling. In: Strategic innovative marketing Springer proceedings in business and economics, pp 391–397
2. West D, Gilpin M, Grant T, Anderson A (2018) “Forrester,” *Forrester*, 26-Jul-2011. [Online]. Available: <https://www.forrester.com/report/WaterScrumFallIsTheRealityOfAgileForMostOrganizationsToday/-/E-RES60109>
3. Chopra R (2018) Software quality assurance: a self-teaching introduction. Mercury Learning and Information, Dulles, VA
4. Rahim MS, Chowdhury AE, Nandi D, Rahman M (2017) Issue starvation in software development: a case study on the redmine issue tracking system dataset. J Telecommun Electronic Comput Eng 9(3):185–189
5. Rahim MS, Hasan M, Chowdhury AE, Das S (2017) Software engineering practices and challenges in Bangladesh: a preliminary survey. J Telecommun Electronic Comput Eng 9(3):163–169

# Multiple Sequence Alignment Algorithm Using Adaptive Evolutionary Clustering



Jyoti Lakhani, Ajay Khunteta, Anupama Chowdhary,  
and Dharmesh Harwani

**Abstract** In the present manuscript, an adaptive evolutionary multiple sequence alignment algorithm is proposed that uses a combination of consensus and SP-score methods. The algorithm searches intermediate pairwise consensus strings that are used to identify the final consensus string for a given set of DNA/RNA/protein sequences. The proposed algorithm is an extension of MPSAGA algorithm that uses positional matrix representation of sequences. An empirical study was performed in the present work to compare the proposed algorithm with the other three contemporary ClustalW, TCOFFEE, and MUSCLE algorithms on the four datasets. The overall observations from the various experiments revealed that the proposed algorithm outperforms than the other algorithms tested in aligning multiple sequences with an average increase of 0.03% in alignment length by inserting 0.02% increased number of gaps.

**Keywords** Multiple sequence alignment · Adaptive evolutionary clustering · Sequence representation · Consensus

---

J. Lakhani · A. Khunteta

Department of Computer Engineering, Poornima University, Jaipur, India  
e-mail: [jyotilakhanimsu@gmail.com](mailto:jyotilakhanimsu@gmail.com)

A. Khunteta

e-mail: [khutetaajay@poornima.org](mailto:khutetaajay@poornima.org)

J. Lakhani

Department of Computer Science, Maharaja Ganga Singh University, Bikaner, India

A. Chowdhary

Department of Computer Science, Keen College, Bikaner, India  
e-mail: [chowdharyanupama@gmail.com](mailto:chowdharyanupama@gmail.com)

D. Harwani (✉)

Department of Microbiology, Maharaja Ganga Singh University, Bikaner, India  
e-mail: [dharmesh@mgsbikaner.ac.in](mailto:dharmesh@mgsbikaner.ac.in)

## 1 Introduction

A multiple sequence alignment is an alignment of three or more DNA or RNA or protein sequences that can organize data in such a way that similar sequence features are aligned together [1, 2]. The goal of the multiple sequence alignment is to reveal features that may be shared by many sequences and to identify alterations that further elucidate functional and phenotypic variability [2]. The main applications of sequence alignment include secondary or tertiary structure prediction, phylogenetic tree construction, function prediction, hidden Markov modeling, PCR primer design, and data validation [2]. The computation of an exact multiple sequence alignment (MSA) of a large set of sequences is extremely complex and is classified as an NP-complete problem [3]. Multiple sequence alignment provides more information than pairwise sequence alignment because it reveals regions which are conserved within a protein family that have structural and functional importance [1]. Multiple sequence alignment is used for distinctive objectives such as performing similarity search of sequences. The approach is used in classification problems (e.g., to classify members in the protein family, to identify close and distant relationship to infer phylogeny).

		Length of each Sequence									
		1	2	.	.	.	.	.	.	.	n
Number of Sequences N	1	C	G	T	-	-	T	C	T	C	
	.	-	G	T	A	A	T	C	G	C	
	.	G	G	T	A	A	G	C	G	-	
	k	C	G	-	A	-	T	C	-	C	

To explain MSA, let us consider a set of three or more DNA/RNA/protein sequences as depicted above. MSA will aim to align these sequences by introducing gaps in each sequence. For example, if there are k number of sequences of N length, then  $S_i$ ,  $i = 1, 2, \dots, K$  and:

$$S = \left\{ \begin{array}{l} S_1 = (s_{11}, s_{12}, \dots, s_{1N}) \\ S_2 = (s_{21}, s_{22}, \dots, s_{2N}) \\ S_3 = (s_{k1}, s_{k2}, \dots, s_{kN}) \end{array} \right\} \quad (1)$$

Consequently, MSA of S will be obtained by inserting gaps ('-') into the sequences in such a way that all resulting sequences  $S_i^*$  will have equal length N and  $S_i^* = S_i$  after removal of all gaps from  $S_i^*$ , and no column will consist of gaps. Consider another MSA  $S^*$  that consists of two sequences  $s_p^*$  and  $s_q^*$  in the alignment. Let us consider two nucleotides  $a$  and  $b$  in the aligned sequence here the score of the sequence alignment will be defined as:

$$\text{score}(a, b) = \begin{cases} \text{match score for } a \text{ and } b & \text{if } a \text{ and } b \text{ are residue} \\ -d & \text{if } a \text{ or } b \text{ are gap} \\ 0 & \text{if } a \text{ and } b \text{ both are gaps} \end{cases} \quad (2)$$

To find the divergence  $d$  of a given set of aligned sequences, the following three methods are used. The divergence between sequences can also be called as the total distance between sequences or the alignment score.

**Consensus Method:** In the consensus method, a common character from each column is searched and the string created in this way is called the consensus string. The total distance between two sequences is calculated by adding a penalty for each character in its column that differs in the sequence from the consensus string. Let us consider  $S$  as a set of sequence wherein  $S = \{S_1, S_2, \dots, S_k\}$  and:

$$S = \left\{ \begin{array}{l} S_1 = (s_{11}, s_{12}, \dots, s_{1N}) \\ S_2 = (s_{21}, s_{22}, \dots, s_{2N}) \\ S_3 = (s_{k1}, s_{k2}, \dots, s_{kN}) \end{array} \right\} \rightarrow S^* \quad (3)$$

$$\text{dist}_i = \sum_{i=1}^k S^* - S_i \quad (4)$$

Here,  $S^*$  is the consensus string of  $S$  and  $\text{dist}_i$  is the distance of  $i$ -th sequence from  $S^*$ .

**Evolutionary Tree Method:** A weighted evolutionary tree is created using sequences where adjacent nodes correspond to the sequence pair. The weight of the tree is defined as the summation of the number of changes between two adjacent nodes in the tree.

**Sum of Pairs (SP score):** The sum of pairs score is the pairwise distance between all sequence pairs. SP score is widely used similarity function. SP score for the two protein sequences is given as predefined BLOSUM or PAM matrix but for more than two sequences, and since the number of possible combinations is too large, SP score needs to be calculated. Let us consider  $S$  as a given set of sequence:

$$S = \left\{ \begin{array}{l} S_1 = (s_{11}, s_{12}, \dots, s_{1N}) \\ S_2 = (s_{21}, s_{22}, \dots, s_{2N}) \\ S_3 = (s_{k1}, s_{k2}, \dots, s_{kN}) \end{array} \right\} \quad (5)$$

$$\text{SP Score}(S) = \sum_{1 \leq i \leq j \leq k} \text{align\_Score}(S_i, S_j) \quad (6)$$

Here, the align score is the alignment score between  $S_i$  and  $S_j$  sequences. The align\_score is equal to the sum of the similarity score of every pair minus gap penalties [4]. The problem of finding a multiple sequence alignment that maximizes the SP score (or minimizes the SP distance) is known to be NP hard [5, 6].

## 2 Literature Review

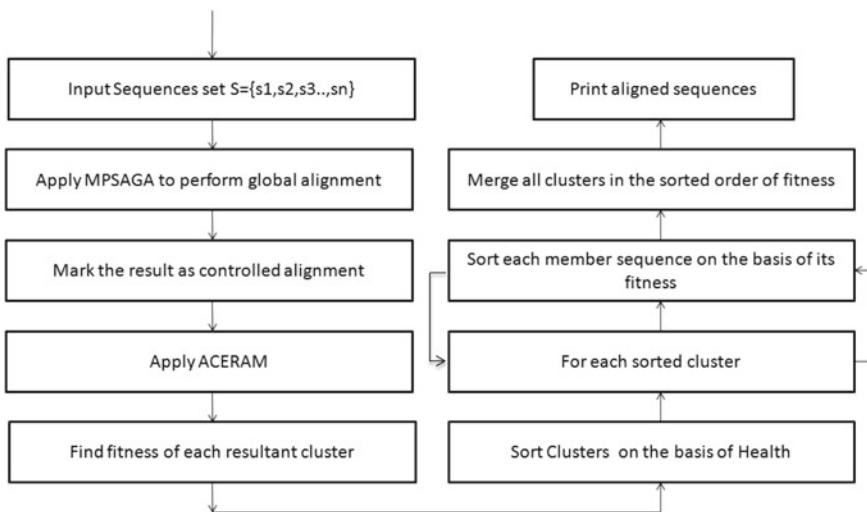
To perform multiple sequence alignment, four distinctive approaches have been discussed in the literature namely global optimization, approximation, heuristic, and probabilistic methods. The probabilistic approach finds the probability of mutation and indels that leads to generating information related to the probability of evolution. Probabilistic methods work efficiently for phylogenetic analysis [7–9]. The global optimization, approximation, and heuristic methods find the optimized score for multiple sequence alignment and are suitable for classification problems. Dynamic programming is a global optimization approach, but the limitation of dynamic programming is that it takes exponential time. Simulated annealing and genetic algorithms have been also used by some researchers to get optimal results [10, 11]. Another approach to overcome the limitation of dynamic programming is to use different search methodologies and improve the efficiency of the global optimization [12–14]. These methods work efficiently with the small datasets but for the large datasets, approximation method is highly useful [15, 16]. Heuristics-based algorithms for multiple sequence alignment can be classified into two groups that are progressive heuristics and iterative heuristics-based algorithms. ClustalW [17, 18] and MUSCLE [19, 20] are well known examples of progressive heuristics and iterative heuristic algorithms, respectively. A combination of heuristic and probabilistic methods has been also suggested by few researchers [19, 21, 22, 23, 24, 25]. Other heuristics-based multiple-sequence alignment methods include simulated annealing [26], branch and bound [27], genetic algorithms [28, 29], Tabu search [30], hidden Markov modeling [31], countless agglomerative and progressive alignment [32], etc. Moreover, some other publically available tools for multiple sequence alignment are Clustal-Omega [33], KAlign [34], MAFET [35], Prank [36, 37, 38], TCOFFEE [39–41], ContraAlign [42], Prime [43], and DiAlign [44–46].

## 3 The Proposed Algorithm

The proposed algorithm is dynamic programming-based multiple-sequence alignment algorithm which is an improved version of the already proposed adaptive evolutionary clustering algorithm MPSAGA [47]. The proposed method was executed with a set of sequences  $S = \{s_1, s_2, s_3, \dots, s_n\}$ . The pair of sequences was identified from the sequence set such as paired\_sequences (PS) =  $\{\{s_1, s_2\}, \{s_3, s_4\}, \dots, \{s_{n-1}, s_n\}\}$ . Using MPSAGA algorithm, these sequences were aligned pairwise. The alignment of these paired sequences was denoted as  $A_{ij}$  where  $i$  and  $j$  denoted the index of the aligned sequences ( $s_i$  and  $s_j$ ). The set of all resultant alignments was denoted as  $A^* = \{A_{12}, A_{34}, \dots, A_{pq}\}$ , where  $p = (n - 1)/2$  and  $q = (n - 1)$ , when there was even number of sequences in the alignment. However, if the odd number of sequences were provided, then one sequence remained unpaired and got paired with the first sequence. For example, if there are 6 (even) sequences to be aligned, then the sequence pairs will

be denoted as  $\text{PS} = \{\{s_1, s_2\}, \{s_3, s_4\}, \{s_5, s_6\}\}$ . But if provided 7 (odd) sequences, then the sequence pairs will be denoted as  $\text{PS} = \{\{s_1, s_2\}, \{s_3, s_4\}, \{s_5, s_6\}, \{s_7, s_1\}\}$  wherein the last unpaired sequence will be paired with the first sequence. The distance between these sequences will be reflected as match\_score. The match\_score of alignment can be calculated by the following formula: Match\_Score = matches reward - mismatches penalty - gap opening penalty - gap extension penalty—indels penalty (7).

The default values used for the parameters in this algorithm are Match\_Reward = +2, Gap\_Opening = -1, Gap\_Extension = -2, Mismatch = -2, and Indel = -2. In the next step, to group similar data items, the resultant pairs were clustered with an adaptive evolutionary clustering algorithm [48]. The step is helpful for the large datasets and can be skipped if the method is applied to the small datasets. The fitness of the clusters is calculated based on the fitness score of the individual clusters, i.e., match\_score, and the clusters are sorted based on their average health [48]. Intracluster sorting is performed with each cluster based on their fitness. Finally, all the clusters are merged and sorted. These aligned sequences resulting from the multiple sequence alignment are sorted according to their match score. The flowchart of the proposed algorithm has been provided in Fig. 1.



**Fig. 1** The proposed MSA-MPSAGA (MPS) algorithm

## 4 Materials and Methods

The present research study was performed on a Windows-based system having an intel i5 processor with 8 GB RAM and 1 TB hard disk. The algorithm was implemented in Java 8 and executed for multiple sequence alignment on nine randomly chosen sequences downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). The NCBI data are publically available for research use, and one can retrieve it by simply submitting the sequence ID.

### 4.1 Datasets Used

To check the performance and accuracy of the proposed multiple sequence aligner, an empirical study was performed in which the following four datasets were used: BAliBASE [49], MattBench [50], Homstrad [51], and Sisyphus [52]. These datasets contained 4–25 sequences. Random sampling was performed on the datasets to create an artificial dataset of 115 sequences as dataset 1, dataset 2, dataset 3, and dataset 4.

### 4.2 Evaluation Criteria

To check the accuracy of the alignments, FastSP v. 1.6.0 [53, 54] was used. FastSP calculates the alignment accuracy with respect to SP score. The accuracy measures provide a value between 0.0 and 1.0. The value of SP score 1.0 indicates the perfect accuracy, and the value of SP score 0.0 indicative of the sequence alignment is incorrect. FastSP also indicates an expansion ratio which is the ratio between the number of matches in the estimated alignment and the observed alignment. The value of expansion ratio less than 1.0 is an indication of over alignment, and value more than 1.0 corresponds to under alignment.

## 5 Results

The proposed algorithm was executed in a single run to perform multiple sequence alignment for the nine sequences downloaded from NCBI (Table 1). The results of multiple sequence alignment have been shown in a similarity matrix. The percent similarity of each sequence with the other sequence is called conservancy, and in the present study, it was calculated using MSA-MPSAGA (MPS) (Table 2). MSA was also performed using ClustalW (CW) [55], TCOFFEE (TC) [56], and MUSCLE (ML) aligners [57, 58]. To compare the results of MSA obtained using all the algorithms tested, visualization method was used. Consequently, the overall results of

**Table 1** Sequences downloaded from NCBI used for empirical study for multiple sequence alignment

S. No.	Sequence ID
1	NM_116010.1
2	DJ399337.1
3	NM_001333948.1
4	BD107596.1
5	MA256607.1
6	NZ_AEC02000093.1
7	NFSD01000006.1
8	HM065552.1
9	NZ_QMBM01000037.1

the multiple sequence alignment were used to construct phylogenetic trees using Phylogenetic Tree Viewer—ETE Toolkit (Table 3).

The empirical study was performed on the four data subsets. The summary of the results is shown in Table 4. The average number of the aligned sequences was observed to be 13, 7, 8, and 10 for the dataset 1, dataset 2, dataset 3, and dataset 4, respectively. The average length of the sequences in the dataset 1 was found to be 765 in which 38 gaps were inserted by the proposed algorithm to align the sequences. The average gap length in the aligned sequences in the dataset 1 was observed to be 9. In dataset 2, the average length of sequences was 260. To align these sequences, average 17 gaps were inserted with an average gap length of 4. A total of 8 sequences of average 421 lengths were aligned by inserting a total of 47 gaps and with an average of 3 basepair long gap length. While ten sequences with average 185 lengths were aligned by inserting 25 gaps with an average of 6 gap length.

The comparison of modeler score and SP score for the four tested algorithms is given in Table 5. It indicated that the MPS algorithm provides the SP score similar to the expected score. The modeler score and SP score of CW, TC, ML, and MPS were observed to be (0.70 and 0.50), (0.78 and 0.77), (0.678 and 0.69), and (0.72 and 0.72), respectively. Each dataset used in these experiments had at most 25 sequences. A total of 115 sequences from a subset of four datasets were used (46 from dataset 1, 36 from dataset 2, 18 from dataset 3, and 15 from dataset 4) (Fig. 2).

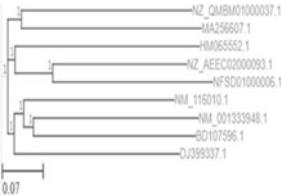
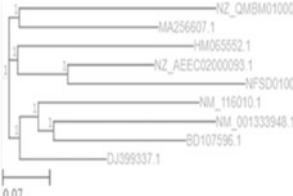
In the other experiment, ten sequence sets of different protein categories were aligned using CW, ML, TC, and MPS algorithms. The consolidated results of the aligned sequence such as average aligned sequence length, the average number of matches in the aligned sequences, number of gaps inserted to align the sequences, and the average gap length inserted in the aligned sequences have been detailed out in Table 6. A comparison of the average length of the aligned sequences for each category of proteins is given in Fig. 3.

Multiple sequence alignment using CW, TC, ML, and MPS aligners provided the average alignment length to be 224.8, 216.8, 229.5, and 230.1, respectively. The proposed algorithm MPS aligned sequences with an increased length of 0.004%, 0.067%, and 0.027% than the CW, TC, and ML algorithms, respectively. The number

**Table 2** Percent conservancy of the nine sequences calculated by MSA-MPSAGA algorithm

NZ_QMBM01000037.1	100.00	41.16	.35.54	39.25	36.11	44.59	40.70	38.28	34.42
MA256607.1	41.16	100.00	.36.76	37.23	34.05	40.95	37.67	34.09	37.40
HM065552.1	35.54	36.76	100.00	39.74	36.51	38.45	40.05	35.64	38.53
NZ_AEEC02000093.1	39.25	37.23	.39.74	100.00	50.66	45.45	43.36	37.39	35.90
NFSD01000006.1	36.11	34.05	.36.51	50.66	100.00	35.49	35.71	34.63	35.37
NM_116010.1	44.59	40.95	38.45	.45.45	44.62	100.00	44.62	46.19	45.34
DJ399337.1	40.70	37.67	40.05	43.36	.35.71	44.62	100.00	41.91	43.52
NM_001333948.1	38.28	34.09	35.64	37.39	34.63	46.19	41.91	100.00	45.11
BD107596.1	34.42	37.40	38.53	35.90	.35.37	45.34	43.52	45.11	100.00

**Table 3** Comparison of phylogenetic trees constructed from multiple sequence alignment of the nine sequences using ClustalW, TCOFFEE, MUSCLE, and MSA-MPSAGA

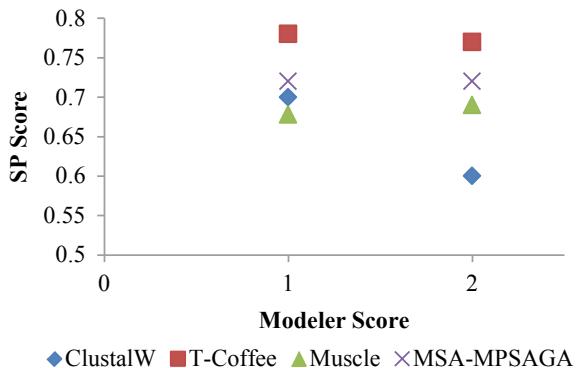
Algorithm	Alignment	Phylogenetic Tree
ClustalW	((NM_116010.1:0.37138, DJ399337.1:0.37811) :0.00541, (NM_001333948.1:0.36473, BD107596.1:0.35478) :0.01468) :0.00541, (MA256607.1:0.38799, ((NZ_AECC02000093.1:0.33187, NFSD01000006.1:0.36210) :0.03898, HM065552.1:0.38383) :0.01235) :0.00235, NZ_QMBM01000037.1:0.38355)	
T-COFFEE	((NZ_QMBM01000037.1:0.28612, MA256607.1:0.30230) :0.02166, (HM065552.1:0.31002, (NZ_AECC02000093.1:0.22498, NFSD01000006.1:0.26847) :0.06202) :0.01069, ((NM_116010.1:0.25288, (NM_001333948.1:0.27673, BD107596.1:0.27221) :0.01500) :0.01454, DJ399337.1:0.27763) :0.00950)	
MUSCLE	((NZ_QMBM01000037.1:0.28612, MA256607.1:0.30230) :0.02166, (HM065552.1:0.31002, (NZ_AECC02000093.1:0.22498, NFSD01000006.1:0.26847) :0.06202) :0.01069, ((NM_116010.1:0.25288, (NM_001333948.1:0.27673, BD107596.1:0.27221) :0.01500) :0.01454, DJ399337.1:0.27763) :0.00950)	
MSA-MPSAGA	((NZ_QMBM01000037.1:0.28453, MA256607.1:0.20120) :0.01344, (HM065552.1:0.25432, (NZ_AECC02000093.1:0.12452, NFSD01000006.1:0.25645) :0.07110) :0.01123, ((NM_116010.1:0.23281, (NM_001333948.1:0.23412, BD107596.1:0.19121) :0.0210) :0.02523, DJ399337.1:0.12532) :0.01425)	

**Table 4** The four datasets analyzed under multiple sequence alignment

Dataset	Avg. no. of seqs	Alignment length	Gaps	Avg. gap length
1	13	765	38	9
2	7	260	17	4
3	8	421	47	3
4	10	185	25	6

**Table 5** Comparison of modeler scores and SP score between the tested algorithms

	CW	TC	ML	MPS
Modeler score	0.70	0.78	0.678	0.72
SP score	0.50	0.77	0.69	0.72

**Fig. 2** The comparison of modeler score and SP score between four benchmarking datasets

of matches in the aligned sequences is shown in Fig. 4. The average number of matches in the aligned sequences was observed to be 29.1, 25.2, 27.7, and 29.2 for CW, TC, ML, and MPS algorithms, respectively. MPS algorithm was found to align the sequences with an increased match of 0.025%, 0.181%, and 0.088% than the CW, TC, and ML algorithms. Comparison based on the number of gaps inserted in the aligned sequences by four multiple sequence alignment algorithms is shown in Fig. 5. The numbers of gaps inserted in the aligned sequences by CW, TC, ML, and MPS aligners were observed to be 271, 280, 280, and 279, respectively. The proposed MPS aligner inserted an increased number of gapes in the aligned sequences than the CW (0.03%), TC (0.014%), and ML (0.014%) algorithms.

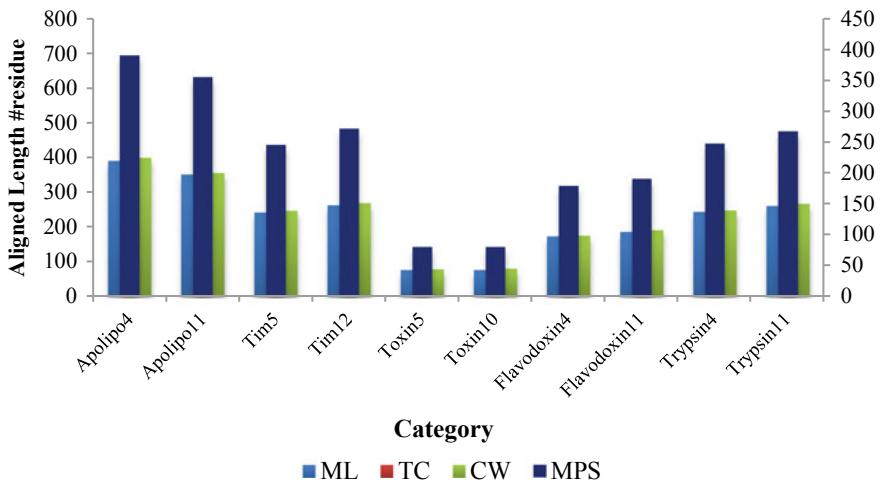
A comparative study based on the match scores of the multiple alignments was also performed using CW, TC, ML, and MPS algorithms. The match score was calculated for the multiple sequence alignments performed by CW, TC, ML, and MPS algorithms that were observed to be 4484.6, 4084.7, 4589.6, and 4682.3, respectively (Fig. 6).

## 6 Conclusion

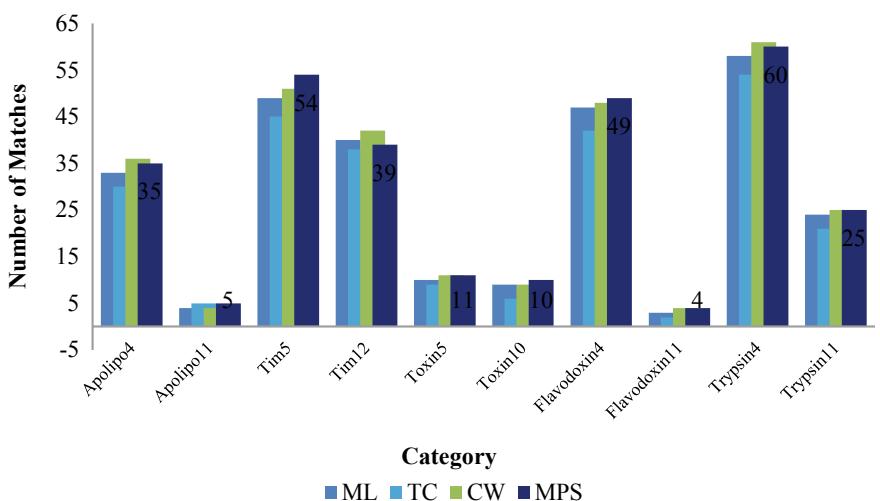
Conclusively, it can be stated that the proposed multiple sequence aligner based on adaptive evolutionary clustering algorithm (MSAMPSAGA or MPS) accurately identifies the sequence alignments. Furthermore, an average increase in sequence alignment length using the proposed aligner was observed to be 0.03% than the other tested algorithms ClustalW, TCOFFEE, and MUSCLE. The phylogenetic trees

**Table 6** A comparative study of the artificial dataset using MSA-MPSAGA with ClustalW, TCOFFEE, and MUSCLE aligners

Category	Num	Avg len. (round)	%Sim	Aligned length						# Match						# Gaps	Match score		
				CW	TC	ML	MPS	CW	TC	ML	MPS	CW	TC	ML	MPS	CW	TC		
Apolipo	4	297	9	398	372	389	382	36	30	33	32	501	528	536	538	2567	2234	2399	2383
Apolipo	11	302	1	354	348	350	351	4	5	4	4	987	998	992	996	9456	8765	9237	9241
Tim	5	247	19	245	237	240	239	51	45	49	47	49	53	50	52	2789	2467	2698	2607
Tim	12	250	15	267	254	261	258	42	38	40	39	276	289	281	285	6753	6299	7569	7578
Toxin	5	67	13	78	76	75	11	9	10	9	70	72	71	70	630	634	651	650	
Toxin	10	67	9	80	68	76	74	9	6	9	8	126	129	127	128	3678	3543	3665	3668
Flavodoxin	4	173	27	173	156	171	171	48	42	47	45	39	41	41	40	2342	2231	2337	2320
Flavodoxin	11	165	3	189	174	184	180	4	2	3	3	281	286	280	282	5679	5349	5663	5621
Trypsin	4	309	21	246	232	242	245	61	54	58	58	59	65	62	63	1987	1785	2894	2889
Trypsin	11	247	8	265	251	259	555	25	21	24	23	325	341	334	335	8965	7540	8783	8857

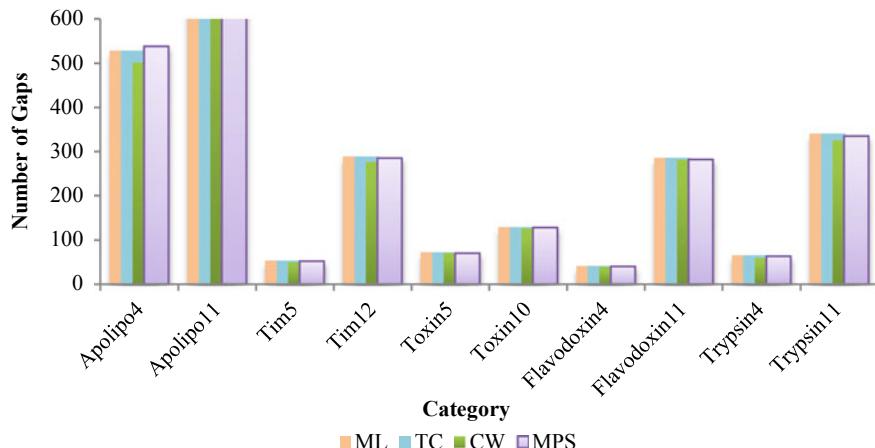


**Fig. 3** Comparison of alignment length of MSA-MPSAGA with ClustalW, MUSCLE, and TCOFFEE multiple sequence aligners

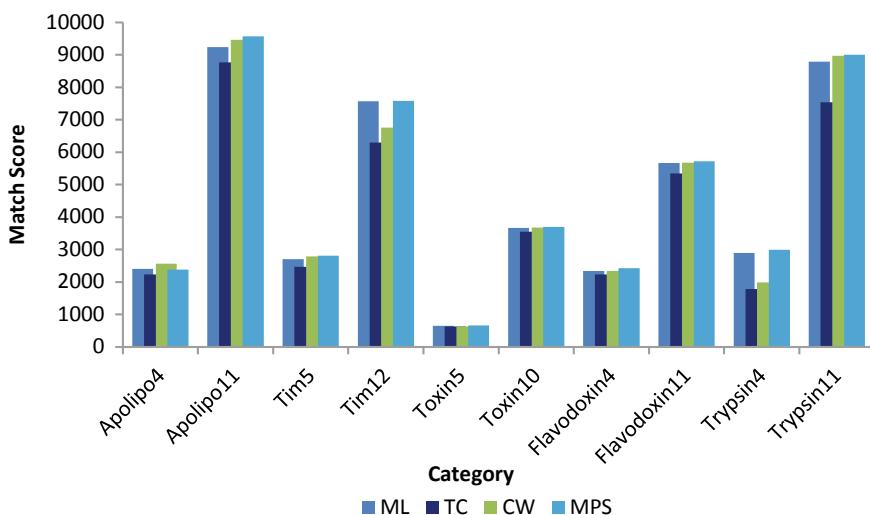


**Fig. 4** Comparison of the number of matches occurred in the aligned sequences using ClustalW, TCOFFEE, and MUSCLE aligners with MSA-MPSAGA

constructed from the MSA obtained from the aligners also indicated that the MPS provides more accurate results. The overall comparison of MPS with the other three tested algorithms showed that the qualitative and quantitative performance of the proposed algorithm is at par as compared to the other aligners. The only limitation of the proposed MPS algorithm is that the algorithm is more useful in doing MSA



**Fig. 5** Comparison of gaps inserted in the aligned sequences by MSA-MPSAGA with ClustalW, TCOFFEE, and MUSCLE aligners



**Fig. 6** Comparison of Match Score for MSA performed by MSA-MPSAGA with ClustalW, TCOFFEE, and MUSCLE aligners

of biological sequences. The implementation of the proposed algorithm in aligning other types of sequences in the varied dataset is a scope of future study.

## References

1. Wiltgen M (2018) Algorithms for structure comparison and analysis: homology modelling of proteins. *Encyclopedia Bioinform Comput Biol: ABC Bioinform* 21:38
2. Carsten K, Notredame C (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25:2455–2465 (Oxford, England)
3. Wang L, Jiang T (1994) On the complexity of multiple sequence alignment. *J Comput Biol* 1(4):337–348
4. Sung WK. Algorithms in bioinformatics: a practical introduction by (CHAPMAN & HALL/CRC mathematical and computational biology series) ISBN 978-1-4200-7033-0
5. Just W (2001) Computational complexity of multiple sequence alignment with SP-score. *J Comput Biol* 8(6):615–623
6. Wang L, Jiang T (1994) On the complexity of multiple sequence alignment. *J Comput Biol* 1(4):337–348
7. Holmes I (2003) Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics* 19(Suppl 1):i147–i157
8. Holmes I, Bruno WJ (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* 17(9):803–820
9. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 26(1):320–322
10. Kim J, Pramanik S, Chung MJ (1994) Multiple sequence alignment using simulated annealing. *Comput Appl Biosci* 10(4):419–426
11. Notredame C, Higgins DG (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res* 24(8):1515–1524
12. Gupta SK, Kececioglu JD, Schaffer AA (1995) Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *J Comput Mol Cell Biol* 2(3):459–472
13. Lipman DJ, Altschul SF, Kececioglu JD (1989) A tool for multiple sequence alignment. *Proc. Natl Acad Sci USA* 86(12):4412–4415
14. Stoye J, Moulton V, Dress AW (1997) DCA: An efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *CABIOS* 13(6):625–626
15. Gusfield D (1993) Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bull Math Biol* 5(1):141–154
16. Pevzner P (1992) Multiple alignment, communication cost, and graph matching. *SIAM J Appl Math* 52(6):1763–1779
17. Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 3(1):237–244
18. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
19. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113
20. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797
21. Loyerntoja A, Milinkovitch MC (2003) A hidden Markov model for progressive multiple alignment. *Bioinformatics* 19(12):1505–1513
22. Edgar RC, Sjölander K (2004) COACH: profile-profile alignment of protein families using hidden markov models. *Bioinformatics* 20(8):1309–1318
23. Edgar RC, Sjölander K (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics* 19(11):1404–1411
24. Loyerntoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 102(30):10557–10562
25. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 15(2):330–340

26. Abhiman S, Daub CO, Sonnhammer EL (2006) Prediction of function divergence in protein families using the substitution rate variation parameter alpha. *Mol Biol Evol* 23(7):1406–1413
27. Reinert K et al (1997) A branch-and-cut algorithm for multiple sequence alignment. In: Santa Fe NM (ed) Recomb97. ACM Press, pp 241–249
28. Gondro C, Kinghorn BP (2007) A simple genetic algorithm for multiple sequence alignment. *Genet Mol Res* 6(4):964–982
29. Notredame C, Higgins DG (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res* 24(8):1515–1524
30. Riaz T, Yi W, Li KB (2005) A tabu search algorithm for post-processing multiple sequence alignment. *J Bioinformatics Comput Biol* 3(01):145–156
31. Rawlings CJ (1995) ISMB-95: Proceedings, third international conference on intelligent systems for molecular biology. AAAI Press
32. Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J Mol Evol* 20(2):175–186
33. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7(1):539
34. Lassmann T, Sonnhammer ELL (2005) Automatic assessment of alignment quality. *Nucleic Acids Res* 33(22):7120–7128
35. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30(14):3059–3066
36. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320(5883):1632–1635
37. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 102(30):10557–10562
38. Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1):205–217
39. Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol* 3(8):1405–1408
40. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* 340(2):385–395
41. Do CB, Gross SS, Batzoglou S (2006) Constrained discriminative training for protein sequence alignment. In: Research in computational molecular biology: 10th annual international conference, RECOMB 2006, Venice, Italy. Springer, Heidelberg, pp 160–174
42. Yamada S, Gotoh O, Yamana H (2006) Improvement in accuracy of multiple sequence alignment using novel group-to-group sequence alignment algorithm with piecewise linear gap cost. *BMC Bioinformatics* 7:524
43. Golubchik T, Wise MJ, Eastoe S, Jermiin LS (2007) Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol Biol Evol* 24(11):2433–2442
44. Morgenstern B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* 15(3):211–218
45. Pei J, Grishin NV (2006) MUMMALS: multiple sequence alignment improved by using hidden markov models with local structural information. *Nucleic Acids Res* 34(16):4364–4374
46. Mirarab S, Warnow T (2011) FASTSP: Linear time calculation of alignment accuracy. *Bioinformatics* 27(23):3250–3258
47. Lakhani J, Khunteta A, Choudhary A, Harwani D (2019) MPSAGA: a matrix-based pairwise sequence alignment algorithm for global alignment with position based sequence representation. *Sādhanā* 44(7):171
48. Lakhani J, Khunteta A, Chowdhary A, Harwani D (2016) Auto-evolving clusters based on rejection and migration. In: Bishnoi SK, Kuri M, Goar V (eds) Proceedings of the International Conference on Advances in Information Communication Technology & Computing (AICTC '16). ACM, New York, NY, USA, Article 98

49. Thomson JD, Plewniak F, Poch O (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* 15(1):87–88
50. Daniels NM, Kumar A, Cowen LJ, Menke M (2012) Touring protein space with Matt. *IEEE/ACM Trans Comput Biol Bioinform* 9:286–293
51. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539
52. Andreeva A, Prlić A, Hubbard TJP, Alexey GM (2007) SISYPHUS—structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res* 35:D253–D259
53. Tang X, Wong DF (2001) FAST-SP: a fast algorithm for block placement based on sequence pair. In: Proceedings of the 2001 Asia and South Pacific design automation conference. ACM, pp 521–526
54. Mirarab S, Warnow T (2011) FastSP: linear time calculation of alignment accuracy. *Bioinformatics* 27(23):3250–3258
55. Thompson JD, Gibson TJ, Higgins DG (2003) Multiple sequence alignment using ClustalW and ClustalX. *Curr Proto Bioinfo* 1:2–3
56. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1):205–217
57. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797
58. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1):113

# Theft Security System for Automatic Teller Machines Using IoT



Vinay Verma, Anjali Verma, Gaurav Sharma, and Anand Sharma

**Abstract** This research paper suggests a system that efficiently and effectively provides a mechanism of anti-theft ATM by using Internet of things and fog computing by considering two solution: First, a instant solution at the beginning of stealing using fog computing and second, cloud-based messages technique system for ATM security and preventing from being stolen by thieves and unsocial elements. As we know that technology reaches its successful step when it fulfills every section of the public or society. As nowadays, it is very common in India that the entire ATM with public money is being stolen by some unsocial elements or thieves. So, this paper proposes an ATM safety while it is being tempered and gives a new system that works using the Internet of things and fog computing methods that would make the ATM and its places very secure and establishes entirely new technologies in ATMs which provides a Theft proof system. The system gives an instant solution while the machine system is being tempered and when criminals try to steal it. Once the ATM experience any vibration, the preventive mechanism will activate and the nearby police station will receive notification messages of the location of the ATM using the cloud. The text message comprises of GPS location of ATMs and also a cautionary message. The fog computing method activates to close the outer shutter of the ATM and the microcontroller to activate the ATM in a new mode of theft preventing by switching on an inbuilt backup supply of the ATM. The remarkable advantage of this system provides the banking system a new way of anti-theft management for money and ATM as well.

---

V. Verma (✉) · A. Sharma

Mody University of Science and Technology, Lakshmangarh, Sikar, India

e-mail: [ervinayv@gmail.com](mailto:ervinayv@gmail.com)

A. Sharma

e-mail: [anand\\_glee@yahoo.co.in](mailto:anand_glee@yahoo.co.in)

A. Verma

Banasthali Vidyapeeth, Banasthali, Niwai, Tonk, India

e-mail: [anjali.professional@gmail.com](mailto:anjali.professional@gmail.com)

G. Sharma

Jaipur National University, Jaipur, India

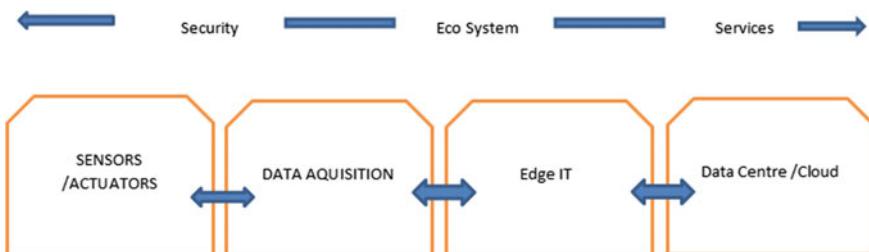
e-mail: [gaurav1981sharma@yahoo.co.in](mailto:gaurav1981sharma@yahoo.co.in)

**Keywords** Internet of things · Cloud computing · Fog computing · GPS tracking system · GSM · Theft prevention ATM

## 1 Introduction

The Internet of things (IoT) directly indicates the usage of wisely coupled system equipment and systems to influence figures collected by embed actuators and sensors in machines and other substantial objects. It uses networks, where things or objects can interact with each other without or minimal human intervention. Equipment-to-equipment solutions is a division of the Internet of things—already use wireless networks to unite equipment together and with the Internet, with the least user involvement, for conveying utilities that assemble the needs of a large range of industries [1]. It empowers objects to communicate with each other and the user. The Internet of things shall surge such choice of utilities, each one needing variable stages of mobility, and potentiality and transmission capacity. Let consider an example, utilities that are linked with public protection or else individual security will usually need less potentiality, but not huge transmission capacity per se. Instead, utilities that offer vigilance could also need greater transmission capacity. Because of the distinct phase of service requests, networks of mobiles could need the capability to recognize the service which is making traffic and meet its exact needs [2]. For example, aware utilities connected to public security or individual fitness would need a higher priority matched to metering figures, which is a usual auditing activity. Another major feature of IoT utilities can be the organization of a great number of the same type of equipment and utilizations. Each equipment and utilization executes the same action and conveys figures to a service center at the meanwhile time. Irrespective of the volume of data transferred by every equipment, this one action could cause congestion of the network [3]. Mobile networks need to deliver numerous processes to protect and better use their competences as conveying such IoT utility processes for administering remotely [1] (Fig. 1).

Such equipment and utilizations could permit smart scheduling, which would enable a proper utilization improvement and decrease the exposure of the system to utilization misconduct [4].



**Fig. 1** Basic IoT architecture

## 1.1 Fog Computing

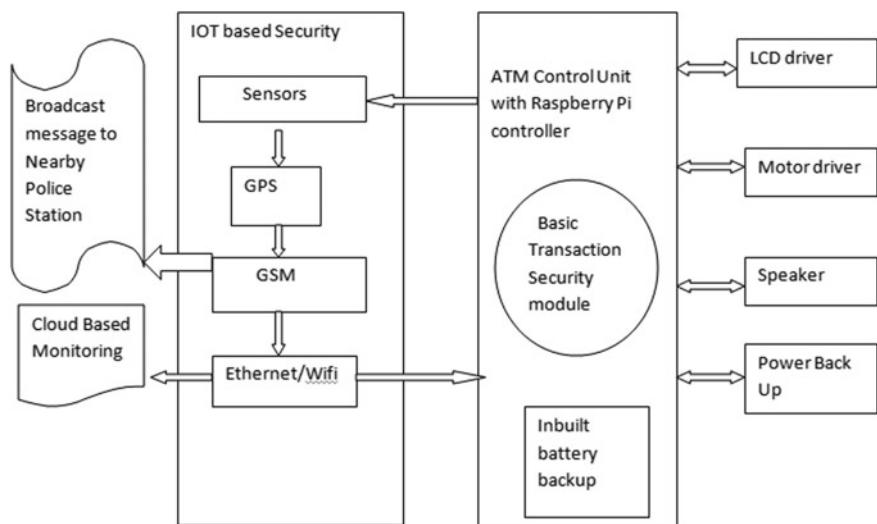
The word “fog computing” or “edge computing” works at the user end instead of working at central cloud systems. It is used to store processes and resources at the edge of the cloud, rather than using frequencies for cloud storage and utilization. Fog computing reduces the necessity for transmission capacity by not transmitting each information bits on the cloud. This type of non-centralized approach drops budgets and enhances productivity. This is an emerging technique of IoT. Fog methods spread the cloud-based computing model to the extent of the network that may not be suitable according to the standard for the cloud because of technical and infrastructure limits.

In fog computing data gathered with the help of sensors and will not be transmitted to the cloud in its place, it is directed to equipment, for example,, network edge, routers, and the access point for dropping the jam due to short transmission capacity. Fog-based methods expand the value of utility and also decrease potentiality.

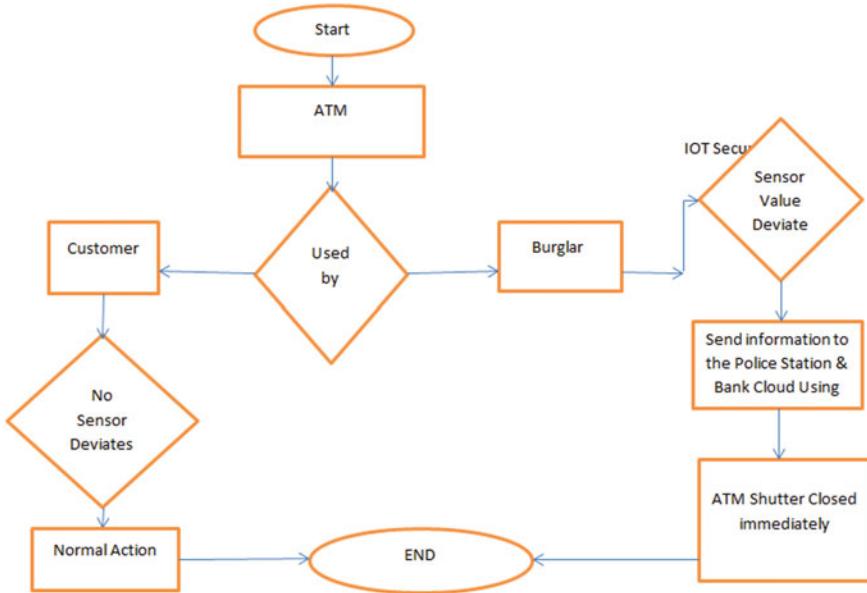
Here are some fundamental methods that fog computing is building the IoT extra secure [5]:

Construction of fog from the earth aimed at security in IoT, presenting a novel phase of shield and outside IT security. Fog computing is reducing the gap between cloud and things. It is enabling the sharing of computing and control, storage, and networking roles nearer to end-user (or “things”). This establishes a completely different ground as compared with IT edge security.

**Computation and Control:** In fog-based computing, supervision control is passed out to end-user as contrasting to be reserved in remotely located data centers or cellular networks. In such a disseminated environment, risk or attacks prerequisite



**Fig. 2** Proposed Model of Internet of things based Security for Automatic Teller Machines



**Fig. 3** Flow chart for working of Algorithm of Internet of things based Security for Automatic Teller Machines

to becoming past fog nodes, which can rapidly identify uncommon activity and be diminished earlier they are passed through to the system.

**Storage of Data:** The similar is correct of data storage within the fog design. Data gathered by or disseminated to end-user equipment are maintained and managed and protected by fog system. Thus, that data will be well secure than whether stored in the user equipment and more accessible than preserved in remotely located data centers

**Networking communication:** While using fog, networking communication is supported at (or near) the end-user in place of steering all traffic through spine networks. This also offers a secrecy benefit. In certain circumstances, such as distributions that implement the Equipment-to-equipment wireless device standard, fog decreases the probabilities of eavesdropping by comprising communication with systems in next to proximity. Fog could guard the minutest resource controlled equipment. The majority of IoT equipment is short and built to function in environments with nominal resources. These resource controlled equipment, which require slight or no ability to protect them in contradiction of cyber-crime, can act as a team with a disseminated multi-layer network of fog and cloud to attain the required phases of safety [5]. A thin client that exists in the equipment is everything that is desired to identify doubtful data in use. If the customer or the fog node identifies uncommon movement, the data which is in usage or transportation shall be highlighted with limitation. The doubtful data or the uncommon activities can be investigated for safety,

privacy, and availability break. The fog node can achieve the more refined safety mechanism required for security.

As stated above, a disseminated array of fog nodes can propose a more protected atmosphere than separate resource-constrained equipment for the storage of important or delicate data.

Fog can support in safety identifications and software up to date on a great sum of equipment to use in universal IoT atmospheres needing each equipment to link to the cloud to bring up to date its record and software, numerous times a day, is unfeasible. But fog nodes are intended to be a dispersed structure for handling security identifications on huge numbers of equipment founded on their uses and/or ownership concurrently without interruption.

Security status for disseminated systems is an accessible and truthful style using fog. In the IoT world, it is vital to be capable to express, in a truthful way, whether a great number of disseminated equipment and systems are functioning firmly and securely. Every other today's hacks are aimed at direct status messages that create processes that seem normal. Fog has the substructure to sense these types of attacks.

Fog can deliver real-time instance response utilities that empower IoT systems to react to concessions without disturbance of service. This is a principally acute purpose in industries where IoT systems and processes offer the enterprises with aim-critical returns production [6].

## 2 Existing System Used in ATM

### 2.1 How ATM Works

The ATM built using generally two input equipment and four output equipment which are;

Input Equipment: Keypad and Card reader.

In all ATM, ISPs play a central role. This will provide communication between ATM and host processors. Whenever we want to use ATM some specific entry can be done by the user. Then, all the input values are delivered to the host system. After receiving the details, the processor at host side checks the user details at the bank server. Once the details are verified, the agreement code is passed to ATM and the user would be able to transfer or withdraw the money.

### 2.2 ATM Types

At present two types of ATM is in which are used by traditional banking System ATM Types

- ATM based on Leased Line (LL-ATM)
- ATM based on Dial.

### 2.2.1 ATM Based on Dial

ATMs which use dial-up connection are to be linked to the host processor via a regular phone line via a modem as well as a toll-free number, or using an ISP through a local access number which modem usually dials.

### 2.2.2 ATM Based on Leased Line

ATM based on leased line is connected to the host processor using a 4-wire, point-to-point, and dedicated link of telephone line.

## 3 Proposed System Used in ATM with IoT

Theft prevention would become a stroke of luck in this progressive technology aware world. Various theft detection systems existing to hook the thief, which can be further amended. If the theft would be prevented, then there will be no loss. The system is targeted for calculating the performance of an operating system. Here, we recommend the “IoT theft prevention system using Raspberry Pi”. Internet of things (IoT) has been leading the microchip technology era with the cloud utilities ruling electronics product segment. In this system, we use a camera along with raspberry pi alongside with a system equipped with LCD view IR for dark vision. The system is power-driven by a 15 voltage power supply. Her camera is used to send normal motion detected pictures using image processing; then, it will be updated on IoT cloud accordingly for proper auditing of the ATM and its current situation. We use IOT Gecko to build an auditing system over the Internet. Thus, the system provides an innovative approach to theft detection using the Internet of things this is the auditing phase of the system [1].

### 3.1 FPGA

Now, we discuss the second phase of the system for that we should know about FPGA. FPGA is the abbreviation for field-programmable gate array. This is an integrated circuit (IC) that can be encoded and constructed with the embedded system designer in the field once it has been contrived. FPGA is not limited to some pre-saved hardware utility; it is usually used by an embedded system developer due to its high adaptability. FPGA generally uses some pre-defined programming logic for

installing or building hardware functionality. The FPGAs are used and configured using hardware description languages (HDL) like Verilog and VHDL and used for a task-specific integrated circuit (ASIC).

This is the cause why FPGA is called field-programmable, as FPGA may simply be rearranged in the field as and when the user required.

## 4 Working of the Proposed Model

The Internet of things, typically named as IoT, is at present the leading technology nowadays and propose vastly profitable opening. The technology in its mostly comprises taking out data from the outside world such as temperature, noise limits, motion, vibration, water limits, position, and so many things. Congregation this type of data requires sensors at all positioned, and recovering the data and deducing it in such style and in certain uses of the data to grow a supervisor system [7]. We just established a theft aware system utilizing a global position system (GPS) and GSM modem and Raspberry Pi. First, a piezoelectric sensor intelligences the occurrence of theft and directs the outcome to the microcontroller. The ATM has a Raspberry Pi controller installed in it which is furnished with sensors such as a heat sensor vapor sensor, and vibration or shock sensor. These sensors are steady at a pre-determined value at the time of fitting of ATM. During the theft, the value of one of the equipment deviates and notified to a pre-determined number (police station) is spread via GSM. The GPS unit sends the location of the ATM. The static location of the principal emergency police station server is saved in the PROM of ATM. At the time any shock has taken place continuously for 60 s, this is identified and message would be sent to the already saved emergency as well as to the Internet of things cloud, where it can be aired to the official persons so that a quick action can be taken the ATM can be saved and decrease the chances of the damages. Primary features of the proposal comprise instantaneous ATM auditing by directing its data information from time to time at Internet of things cloud and where statistics can be accessed by the online auditing system concerning ATM security and if it flops and theft happens then essential act could be executed. Whenever a theft occurrence take place, vibrating sensor observes and directs the signals to the microcontroller, by the use of GPS specific locations from where theft occurrence is originated [8]. We also used FPGA which is mostly used to trace the position of a nearby PCR and delivers computerized message (Fig. 2).

## 5 Types of Equipment and Tools Used in Proposed System

### 5.1 Theft Finding Unit

All other components like the other sensors and GPS and GSM units are related via ATM control unit. The LCD screen shows tiny text SMS to preserve track of the functioning of the system. The alarm is prompted when a theft is identified. Accelerometer is used to spot smash or rollover of the ATM and directs signals once a theft happens to the microcontroller. The ultrasonic sensor senses the theft happened due to a hindrance or not [7].

### 5.2 Location Finding Unit

GPS—global positioning system unit is used in ATMs together for tracing and map reading. Tracking systems assist a base station to have a pathway of the ATMs without the intervention of the driver where like map reading system supports the police to reach the endpoint. When theft happens in any area, then the GPS tracks the location of the ATM and directs the information to the specific pre-determined police station through GSM. As an additional option, location detection can be done using the Google Maps interface [9, 10, 11].

### 5.3 SMS Unit GSM

Global system for mobile communication unit is used as a means to alert precautionary utilities by using an SMS message.

### 5.4 Inbuilt Alarm

It provides a loud warning for a theft happening.

## 6 Scope of the Study

The scope of the study as follows:

Develop a system to track ATM in real-time. Develop a system to perceive theft and inform on cloud and police station. Thus, this system is very useful and can be

used in all aspects of ATM systems and its security as well as the safety of money [12].

## 7 Algorithm for Theft Detection

At whatever time a theft is happened, vibration or sound sensor and other sensor deviates from its normal value and send the deviation of values as theft occurring scenario and then the shutter of the ATM center will be closed immediately using fog computing technologies which uses end equipment of data before it is being sent to the Internet of things cloud and the changes will be displayed in the display, subsequent this the deviations will be reorganized and will be acknowledged to the online auditing system and using GSM the message will be sent to the nearest police station [13]. The vibration sensor is installed here to detect the ups and downs inside the ATM area. As soon as the burglar uses his driller and reaper for breaking the bolt, then using fog computing and vibration sensor's data and its deviation from its normal values, the ATM area's door shutter will be locked immediately. Next, to this, the SMS will also direct to the particular banks and the nearest police station using GPS-GSM unit to nearby pre-saved police station numbers and broadcast the sensor information on the Internet of things cloud as an instantaneous effect (Fig. 3).

## 8 Conclusion

The progression was made to implement the system in real-time on a battery of an ATM while placing the system inside the ATM such that it is not stolen by the thief. The system developed effectively provides the use of connected equipment or the Internet of things in ATMs. The system uses a combined GPS plus GSM unit for tracking the location of the ATM by the global positioning system's antenna implanted in IoT enabled ATM. Thus, the whole system is a mixing of numerous current communications and embedded technologies. Using it, the system can be affordable as open-source tools. Security standards are preserved by cellular network suppliers so the security of the network is more decent. Thus, the proposed model is very useful and can be used in all aspects of ATM safety.

## References

1. Kim J, Lee J, Yun J (2013) M2M service platforms: survey issues, and enabling technologies. IEEE Communications Society
2. Atzori L, Iera A, Morabito G (2010) The Internet of Things: a survey. In: Computer networks. Elsevier, 31 May 2010

3. Evans D (2011) The Internet of Things-how the next evolution of the internet is changing everything. Cisco Internet Business Solutions Group (IBSG)
4. Stankovic JA (2014) Research directions for the Internet of Things. IEEE Internet Of Things J 1, February 2014
5. Ben-Salem M, Angelos S, Keromytis D (2012) Fog computing: mitigating insider data theft attacks in the cloud. In: IEEE symposium on security and privacy workshop (SPW)
6. Bonomi (2011) Connected vehicles, the internet of things, and fog computing. In: The eighth ACM international workshop on vehicular inter-networking (VANET), Las Vegas, US
7. Kannan P, Vidya M (2013) Design and implementation of a security-based ATM theft auditing system, July 2013
8. SIM808 GPRS/GSM + GPS Shield v1.1
9. Khan A, Mishra R (2012) GPS–GSM based tracking system. Int J Eng Trends Technology 3(2)
10. Ramani R, Valarmathy S, Vanitha NS, Selvaraju S, Thangam R, Thiruppathi M (2013) Vehicle tracking and locking system based on GSM and GPS. IJ Intell Syst Appl 5(9):86–93
11. Maiti S, Vaishnav M, Ingale L, Suryawanshi P (2016) ATM robbery prevention using advance security. Int Res J Eng Technol (IRJET) 03(02)
12. Positive Technologies (2018) ATM logic attacks: scenarios, November 14, 2018. [online]. <https://www.ptsecurity.com/ww-en/analytics/atm-vulnerabilities-2018/>. Accessed on 25 June 2019
13. Prasanth Ganesh GS, Balaji B, Srinivasa Varadhan TA (2011) Anti-theft tracking system for automobiles (AutoGSM). In: IEEE international conference on anti-counterfeiting, security and identification

# Tree-Based Multi-Keyword Rank Search Scheme Supporting Dynamic Update and Verifiability upon Encrypted Cloud Data



Pawan Kumar Tanwar, Ajay Khunteta, Vishal Goar, and Manoj Kuri

**Abstract** Due to large scale use and many applications of cloud maximum data, owners upload the information at cloud space to save time and local disk space. Here, the authors provided a TBMKRS (Tree-Based Multi-Keyword Rank Search) scheme which also supports dynamic update (insert/delete) and verifiability of encrypted information upon cloud. For the evaluation of performance and analysis of result, the Enron data set have been used by the authors.

**Keywords** Multi keyword rank search · Dynamic update · Verifiability etc

## 1 Introduction

The concept of cloud storage has been spread up at very large scale in last few years due to its unique benefits. Everybody wants to keep his or her data at cloud server to shorten the overhead of computing. There are some limitation in this concept, that is security and privacy risks of data, because generally cloud servers are treated as non trusted entities. To deal with these types of risk, data are uploaded at cloud server in the encrypted form. The encrypted has its different types of challenges like searching anything from encrypted data is very tedious[1–11].

SE (Searchable Encryption) schemes take care of these types security risks. A huge amount of research have been done in this area in recent years like single keyword,

---

P. K. Tanwar (✉) · A. Khunteta  
Poornima University, Jaipur, India  
e-mail: [pktbkn@gmail.com](mailto:pktbkn@gmail.com)

A. Khunteta  
e-mail: [khutetaajay@poornima.org](mailto:khutetaajay@poornima.org)

P. K. Tanwar · V. Goar · M. Kuri  
Engineering College Bikaner, Bikaner, India  
e-mail: [dr.vishalgoar@gmail.com](mailto:dr.vishalgoar@gmail.com)

M. Kuri  
e-mail: [kuri.manoj@gmail.com](mailto:kuri.manoj@gmail.com)

multi keyword, conjunctive keyword, and ranked search. Along with these, some other things have also been proposed for research; these are dynamic update and verifiability. Here, the authors provided a TBMKRS (Tree-Based Multi-Keyword Rank Search) scheme, which also supports verifiability and dynamic update.

For improving the efficiency, the authors have also considered the TF X IDF method and the rank relevance score to find out the intermediate results. The analysis of the security has also done against the threat models. To complete the research, the stakeholders of the whole system are being considered. These stakeholders are owner of data, server (cloud), and the user of data. The owners of information upload his information at server (cloud) in a secured way, whereas the user of data queries the data from the server (cloud). For the privacy, safety, security, and integrity of data, various types of measures and protocols have been taken into consideration. The advantage of this research is that time and space for handling the data are saved.

The issue of dynamic update of data is also very significant because once the data owner outsource the data at cloud server but when modifications are required to be done upon data then the whole thing become very cumbersome. First of all, the data owner downloads all the data at local machine and makes modifications and then again uploads the data at cloud server; in this whole process, lot of time and local machine space are waste. Hence, the authors proposed the concept of dynamic update to make the changes in the data dynamically.

There are more than one methods of dynamic update, and one of them is initially keep some blank space in the data repository and later on use these blank spaces to append or insert the additional data. In addition, the space vacated after deleting some data from the existing data can also be used to insert new data [12–14].

## 2 Related Work

First of all, Song [15] provided a method for search with single keyword. The scheme was secure but the cost of searching was very high. Similarly, Goh [16] described the scheme with pseudorandom function and bloom filter. This scheme provided positive-false output. Mitzenmacher and Chang designed 2 schemes with dictionaries. A single keyword search method which supports fuzzy search has been designed by Li et al. [17]. Boneh et al. [18] developed a public key method which supports disjunctive and conjunctive search like range and subset query. Based on inverted index, Wang et al. [19] proposed a public key SE scheme.

Rank search developed to deal with shortcomings of Boolean searching. Wang et al. [20] utilized tf x idf and inverted index to form a symmetric encryption with order preserving. Only single keyword searching is favored by this scheme. First of all, Cao et al. [17] designed a basic multi keyword rank search scheme with inner product computing with reduced overhead. This scheme avoided the various significances of keywords. A scheme supports rank and fuzzy search developed by Fu et al. [21]. This scheme has been processed by stemming algorithm, bloom filter, and LSH. Index structure scheme with MDB tree has been designed by Sun et al.

[22]. To form a index tree with hierarchical cluster by using k means algorithm has been proposed and to provide multi keyword rank search a special KBB index tree has been designed by Chen et al. [23].

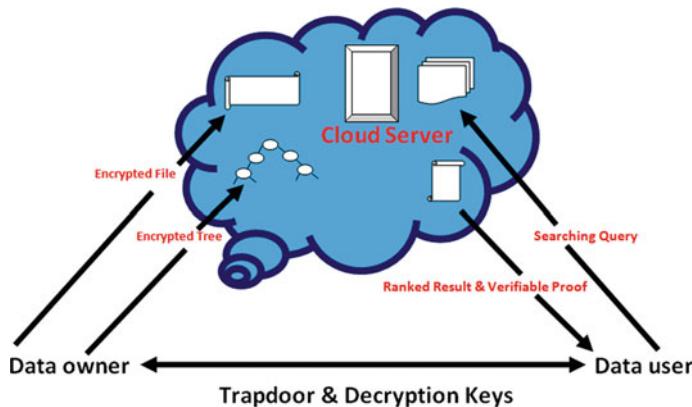
To provide dynamic SSE scheme, Kamara et al. [24] formed a dynamic encrypted index. Further, this scheme has been improved by using KBB tree. By using pseudorandom padding and homomorphism encryption, Wang et al. provided an efficient dynamic index scheme. Bilinear map accumulation tree based update has been applied by Deng and Wan et al. [25]. Symmetric structure encryption method using binary tree has been designed by Chow and Lai [26]. Deng and Wan [25] provided a basis for applying verifiability base upon homomorphism MAC. Merkel hash and MDB tree are joined by Sun et al. [19] to get verifiability.

### 3 Problem Formulation

#### 3.1 System Architecture

The system architecture includes mainly 3 things—server, owner, and user as elaborated in Fig. 1.

**Data Owner**—Firstly, the owner of data enciphers the text files FL by applying some algorithm and forms the encrypted TR (index tree) to enhance the effectiveness of searching. At next step, enciphered files and index tree are uploaded to server (cloud). Further, the secure key will be send to the authorized user. If the owner wants to modify the data at cloud server, at that time the request for update will be directed to the server.



**Fig. 1** System architecture

**Table 1** Notations

FL	Plaintext files of data owner where $FL = \{f_1, f_2, f_3, \dots, f_n\}$
WD	Dictionary of words shared between user and owner of data where $WD = \{wd_1, wd_2, \dots, wd_n\}$
CD	Ciphered file collection saved at cloud server where $CD = \{cd_1, cd_2, \dots, cd_n\}$
TR	Non ciphered index tree formed from CD
IT	Ciphered index tree formed from TR
QR	Query vector initiated by user includes n words in WD where $QD = \{qd_1, qd_2, \dots, qd_n\}$
TDR	Trapdoor formed by the QR and uploaded to the server
RK	Output from server in the form of top k ciphered files
PF	Plaintext files output from decryption of RK

**Cloud Server**—The server keeps all the data outsourced by the owner. The encrypted index tree and the encrypted files sent by the owner are saved at cloud server. After getting the query for search from the user, the server traverses in the index tree to get the required files and outputs the top k ranked documents to the user. Along with it, the cloud server also sends the proof of verifiability to the user also. After receiving the request for update from the owner, the server updates the files as well as index tree.

**Data User**—The user of data uploads a query for searching the keywords. It uses a key to form a trapdoor and send to cloud. After getting the result, the user of data decipher the result with the key shared by the data owner. The returned verifiable proof is used to verify correctness and integrity of output.

### 3.2 Notation

Notations used are described in Table 1.

### 3.3 The Proposed Scheme

**Tree Based Index Formation**—The authors have formed the index on the basis of binary tree following the Xia's scheme [19]. This tree-based index will enhance the searching efficiency at large scale. In this formation, the every vertex v in the tree is defines as— $v = \{ID, Al, Ar, I, h\}$

Here, ID = vertex identity,  $Al$  = pointer to left child vertex,  $Ar$  = pointer to right child vertex,  $I$  = index vector,  $h$  = hash value of vertex v for verifiability. Every leaf vertex is attached to a file. The TreeForm process is shown in Algo.1. The Algo.2. uses

trapdoor and the indexes to compute the similarity for getting the top k documents. The owner of data forms the hash tree on the basis of index tree by applying Algo.3. The proof and minimum hash sub tree mintr is provided by Algo.4. After getting the results and proof, the user of data verifies the result for authorization, correctness and completeness. Algo.5. provides the output from server in the form of top k ciphered files

### **TreeForm (Algo.1.)**

```

INPUT- plaintext files FL, the index vector I
OUTPUT- the index tree TR
FOR every file fli in FL do
    Initialize the leaf vertex v.ID = ID(fli), v.Al = null, v.Ar = null, v.I[i] = TFfli,wdi
    for i ∈ [1, n]
END FOR
WHILE root vertex not formed DO
    Form parent vertex for every 2 vertex v' & v'' v.ID = ID(v),v. Al = v', v.Ar =
    v'',v.I[i] = max{v'.I[i].v''.I[i]}for ie[1,n]
END WHILE
RETURN tree TR

```

### **Search(v) – (Algo.2.)**

```

INPUT - vertex v, index tree, TDR(trapdoor), THR(threshold)
OUTPUT- top k files RK
Calculate relevance score RS = v.I.TDR
IF the vertex v is not a leaf vertex THEN
    IF RS > THR THEN
        Search(v. Al)
        Search(v.Ar)
    END IF
    ELSE
        IF RS > THR THEN
            insert the vertex and score into RK
        IF length(RK) > k then
            Sort RK and erase the output with min score
            THR = min(RK)
        END IF
        END IF
        END IF
    RETURN RK

```

### **Hash Tree Formation (Algo.3.)**

```

FOR every leaf vertex DO
    v.h = hash(v.ID||Φ(fli) //Φ(fli) is the text of file
END FOR
FOR every non leaf vertex DO

```

```

v.h = hash (v.ID || h.A/ || h.Ar)
IF vertex is root vertex THEN //sign
or = sign(v.h || ts) //ts = time stamp
END IF
END FOR

```

#### **Min hash sub\_tree (Algo.4.)**

```

INPUT- Output RK, TR (index tree)
OUTPUT- min hash sub_tree mintr
FOR every vertex v in RK DO
append v in mintr
WHILE v is not root vertex DO
append v's parent vertex and v's brother vertex into mintr
v = v.parent
END WHILE
END FOR
RETURN mintr

```

#### **Verifiability (Algo.5.)**

```

INPUT- min hash sub_tree mintr,
OUTPUT- RK
IF the sign of root vertex is true then //authorized
IF verifiability of every vertex in mintr is true then//authorized
Recomputed the hash val for vertex in RK
IF recalculated val = val in mintr then
Again search mintr applying identical TDR//correct
IF the result of again search = R then//correct & complete
RETURN true
END IF
END IF
END IF
END IF

```

#### **Updt proof (Algo.6.)**

```

INPUT- document fupdt
var = {ins,del,mod}
Encipher the document fupdt to cupdt
IF var = ins then
append the cupdt into leaf vertex
END IF
IF var = del THEN
search & initialize vertex of cupdt = NULL
END IF
IF var = mod THEN
search & update vertex of cupdt

```

```

END IF
again form fresh index tree
form min sub tree tupdt for cupdt by Algorithm4
RETURN{tupdt,var,cupdt}

```

#### **Updt-(Algo.7.)**

```

INPUT- modified document cupdt, sub-tree tupdt, update var
IF var = insert THEN
append the cupdt into the group CD
END IF
IF var = del THEN
search and erase the file
END IF
IF var = mod THEN
search and change the file to cupdt
END IF
change relevant vertex in TR to tupdt

```

### **3.4 Dynamic Update**

The information kept at server could be erased and changed, new files could be inserted, and the scheme should favor dynamic update. Here, 2 things to be considered. One thing is that some blank entries may be kept in the dictionary in advance. This method can take care about most of the entries by accommodating them with the reduced overhead.

The other is the update of document group that will impact encrypted index tree and group of documents. The owner of data keeps safe an index tree (plain text) at local level and provides required data for updating in the enciphered way as shown in Algo.6. After getting the information for update, the server utilizes the information to update relevant vertex in index tree and file *cupdt* to update the set of documents as shown in Algo.7.

## **4 Conclusion and Future Work**

The authors presented here a tree-based multi-keyword rank search scheme which also supports dynamic update and verifiability over encrypted cloud data. The algorithm used here obtains better search efficiency than linear search. Here, the verifiability is the most important thing which returns authenticity, correctness, completeness, and freshness of data. There is few limitation of the scheme. In the proposed scheme, the owner of data is responsible for outsourcing and updating data. Hence, the data owner requires saving the unencrypted index tree and the data which are

**Table 2** Comparative study of this paper with other schemes

Scheme	Construction	Verifiability	Dynamic Update
This paper	Binary Tree	Yes	Yes
[25]	MDB Tree & Interest Model	No	No
[21]	MDB Tree	No	No
[18]	KBB Tree	No	Yes

mandatory to recalculate the IDF values. This type of active owner of data is not suits the cloud model. It might be a meaningful but a tedious future work to develop a dynamic SE scheme whose update operation could be fulfilled by the cloud server itself. The comparative study of the proposed scheme and other schemes are shown in Table 2.

## References

1. Moffat A, Zobel J (1998) Exploring the similarity space, ACM SIGIR, pp 18–34
2. Li H, Hou YT, Lou W, Liu X, Sun W (2015) Catch you if you lie to me: efficient verifiable conjunctive keyword search over large dynamic encrypted cloud data. In: 2015 IEEE conference on computer Comm, pp 2110–2118, April 2015
3. Perrig A, Song D, Wagner D (2000) Practical techniques for searches on encrypted data, In: Proceeding of IEEE symposium on Sec. and Priv., pp 44–55
4. Lou W, Ren K, Cao N, Wang C, Wang Q, Li J (2010) Fuzzy keyword search over encrypted data in cloud computing. In: Conf. on Info. Comm., pp 441–445
5. Shen X, Zhou L, Liang X, Luan T, Yang Y, Li H (2016) Enabling fine-grained multi-keyword search supporting classified sub-dictionaries over encrypted cloud data. IEEE Trans Dependable Secur Comput, 312–325
6. Papamanthou C, Kamara S (2013) Parallel and dynamic searchable symmetric encryption. In: Sadeghi A-R (ed) FC. LNCS, vol 7859, pp 258–274. Springer, Heidelberg
7. Waters B, Staddon J, Golle P (2004) Secure conjunctive keyword search over encrypted data. In: Jakobsson M, Yung M, Zhou J (eds) ACNS 2004, vol 3089. LNCS. Springer, Heidelberg, pp 31–45
8. Huang F, Sun X., Fu Z., Shu J., Ren, K., Enabling personalized search over encrypted outsourced data with efficiency improvement. IEEE Transaction on Parallel Distributed System, 2546–2559, 2016
9. Mitzenmacher M, Chang YC (2005) Privacy preserving keyword searches on remote encrypted data. In: Ioannidis, YA., Keromytis JM (eds) ACNS, LNCS, vol 3531, pp 442–455. Springer, Heidelberg
10. Lou W, Ren K, Li M, Wang C, Cao N (2013) Privacy-preserving multi-keyword ranked search over encrypted cloud data. IEEE Trans Parallel Distrib Syst, 222–233
11. Persiano, G, Ostrovsky R., Crescenzo G., Boneh, D., Public key encryption with keyword search. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004, LNCS, vol. 3027, pp. 506–522. Springer, Heidelberg, 2004
12. Tanwar PK, Goar V, Khunteta A (2018) Design and analysis of search algorithm with B-tree and commutative key RSA for dynamic updation in Cloud Computing. Int J Curr Adv Res 7(7(H)): 14414–14418
13. Tanwar PK, Goar V, Khunteta A (2017) Performance evaluation of multi keyword ranked search schema called BDMRS-CM & EDMRS-BM in cloud computing. Int J Eng Sci 24:42–51

14. Tanwar PK, Goar V, Khunteta A (2016) Design of new multi keyword ranked search scheme and validation for cloud computing. In: Proceedings of AICTC—2016, Bikaner, India
15. Hou YT, Lou W, Song W, Wang B (2015) Inverted index based multi-keyword public-key searchable encryption with strong privacy guarantee. In: Computer communications, pp 2092–2100
16. Goh EJ (2003) Secure indexes Cryptology ePrint Archive, Report 2003/216
17. Lou W, Ren K, Li J, Cao N, Wang C (2010) Secure ranked keyword search over encrypted cloud data. In: Proceedings of the 2010 IEEE 30th international conference on distributed computing systems, pp 253–262
18. Waters B, Boneh D (2007) Conjunctive, subset, and range queries on encrypted data. In: Vadhan SP (ed) TCC LNCS, vol 4392, pp 535–554. Springer, Heidelberg
19. Wang Q, Wang X, Sun X, Xia Z (2016) A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. *IEEE Trans Parallel Distrib Syst* 340–352
20. Zou Q, Lai RW, Chow SS, Du M, Wang Q, He M (2016) Searchable encryption over feature rich data. *IEEE Trans Dependable Secur Comput*
21. Ren K, Sun X, Guan C, Wu X, Fu Z (2017) Towards efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement. *IEEE Trans Inf Forensics Secur*, 2706–2716
22. Sun W et al (2014) Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking. *IEEE Trans Parallel Distrib Syst*, 3025–3035
23. Chen C et al (2016) An efficient privacy-preserving ranked keyword search method. *IEEE Trans Parallel Distrib Syst*, 951–963
24. Roeder T, Papamanthou C, Kamara S (2012) Dynamic searchable symmetric encryption. In: ACM Conf. on Comp. and Comm. Sec., pp 965–976
25. Deng RH, Wan Z (2017) V P search: achieving verifiability for privacy-preserving multi keyword search over encrypted cloud data. *IEEE Trans Dependable Secur Comput*, 99
26. Chow C, Lai RWF et al (2017) Parallel and dynamic structured encryption. In: Deng R, Weng J, Ren K, Yegneswaran V (eds) *Secure communication 2016. LNCS*, vol 198, pp 219–238. Springer, Cham

# Techniques, Applications, and Issues in Mining Large-Scale Text Databases



Sandhya Avasthi, Ritu Chauhan, and Debi P. Acharjya

**Abstract** The discovery of knowledge from large-scale text data or semi-structured data is very difficult. In text mining, useful information is extracted out of such large text corpus which fulfills a user current information need. This process is being exploited by various organizations for quality improvement, business need, and understanding user behavior. The text available in unstructured and semi-structured form can come through sources such as medical, financial, market, scientific, and others documents. Text mining applies quantitative approach to analyze massive amount of textual data and tries to solve information overload problem. The main objective is to review text mining techniques, application areas, and existing issues.

**Keywords** Text · Mining · Information retrieval · Machine learning · Information extraction

## 1 Introduction

Due to immense popularity of computer applications among users, many institution and organization are keeping data in various available formats for future use. These data are available mainly in two formats, structured and unstructured. The text data available through different digital libraries, repositories, sources like blogs, social network, and e-mails are really huge [1]. The task of processing such data to determine useful patterns and extracting valuable knowledge from these data is very difficult [2]. The process to mine such text data using conventional tool could require

---

S. Avasthi · R. Chauhan (✉)  
Amity University, Noida, India  
e-mail: [rituchauha@gmail.com](mailto:rituchauha@gmail.com)

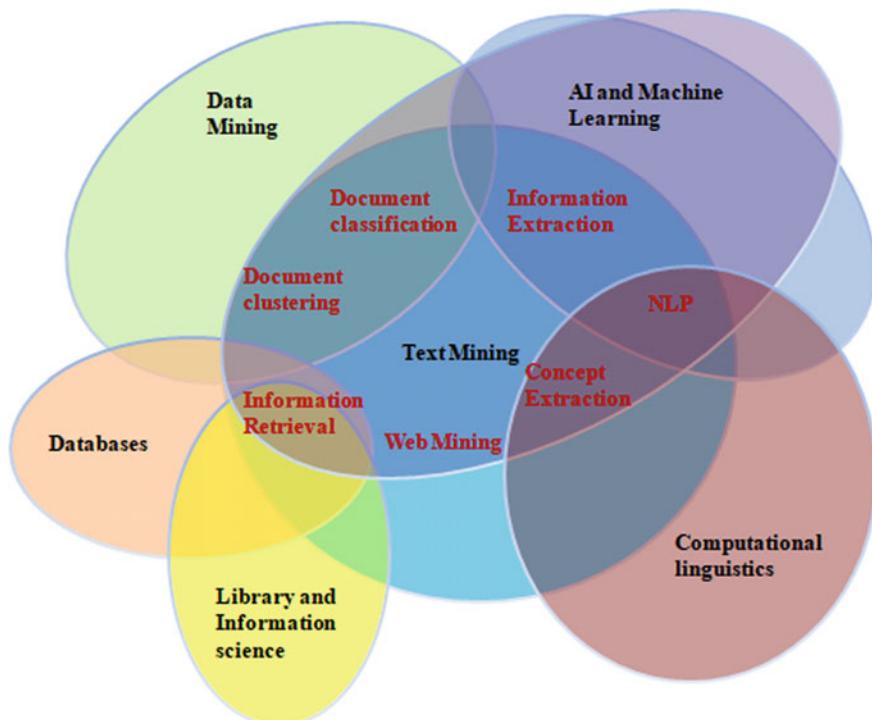
S. Avasthi  
e-mail: [Sandhya\\_avasthi@yahoo.com](mailto:Sandhya_avasthi@yahoo.com)

D. P. Acharjya  
Vellore Institute of Technology, Vellore, India  
e-mail: [dpacharjya@gmail.com](mailto:dpacharjya@gmail.com)

lot of effort. Text mining views any text data of any size as one document. The text mining utilizes techniques from areas like data mining, statistics, machine learning, information retrieval, and computational linguistics [3, 4]. The text mining and relation with other domain is illustrated through Venn diagram in Fig. 1 [4]. Text mining handles natural language text documents that have no proper format by applying techniques like clustering, classification, topic modeling, summarization, and sentiment analysis [5, 6]. Many online applications like search engines, e-mail filters, product recommendation, fraud detection, and social media analytics use text mining extensively for feature extraction, sentiment analysis, and trend analysis [4, 7].

While searching keywords like document classification, information extraction, document clustering, information retrieval, concept extraction, web mining, and NLP on IEEE Explore digital library, the collected data are presented in Table 1. These publications are dated from 1955 to 2019. The maximum number of research items is published in the field of Information Extraction. “Total” gives the number of publications including conferences, chapters, journal article, books, magazines, courses, and standards.

Mining of text data or text documents mainly includes three activities like data preparation, analysis of text corpus, and advanced NLP [8]. The preparation of text



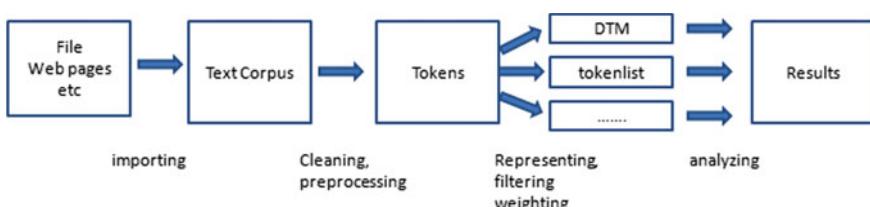
**Fig. 1** Venn diagram showing text mining and other areas

**Table 1** Showing publication count according to keywords as on date July 16, 2019

Keyword	Conferences	Journal	Total
Document classification	4993	353	5481
Information extraction	68,384	11,724	81,514
Document clustering	3022	279	3354
Information retrieval	40,446	7819	49,907
Concept extraction	4619	1080	6839
Web mining	12,455	625	13,451
NLP	25,164	2139	27,987

data completes in five steps; importing text, string operation on text corpus, preprocessing, text representation, and filtering. The most common knowledge representation is document term matrix (DTM). The sequence through which these steps are performed is illustrated in Fig. 2. An important operation in the process is known as parsing that is done by joining, splitting, and extracting strings in texts. Some preparatory steps like tokenization, normalizing features, and removing stop words are known as text corpus preprocessing [8]. When a text is broken down or split into tokens like words or letters, it called tokenization. The most frequently used words like “the,” “are,” and “is” in English are not important for analysis in text mining. Such frequently occurring words in English are called “stop words.” The stop words are sometime removed in preprocessing steps, but in case of prediction model, they become useful in the document so they are kept.

The overview of many existing techniques in text mining field to analyze text corpora efficiently is given in this paper, and problems that occur in the process of text mining is identified. The background into previous work, techniques, application domain, and problems is explained in Sect. 2, Sect. 3, and Sect. 4 respectively. The main problem that exists in text mining is discussed in Sect. 5, and conclusion is given in Sect. 6.

**Fig. 2** Processing steps of text documents

## 2 Background

A typical text mining process starts with collecting text data. Next sequence of steps is extraction, preprocessing, transformation, feature extraction, selection, and evaluation of text corpus [7, 9]. In this paper, text mining applications in various domains are studied and open problems are discussed [6]. Unstructured text is very difficult to process in comparison with structured or tabular data using conventional data mining methods and tools [10]. Also, text mining use in bioinformatics, security system, and business intelligence is explained. Use of natural language processing along with named entity recognition is making task of text mining easier. Entity recognition is recognizing person, person, and location [11]. The database framework called MEDLINE is public repository of biomedical document and utilizes techniques like NER, text classification, abbreviation, testing, and relationship extraction to perform various tasks [12, 13]. This framework works by removing details and extracting only relevant text information, and thus, it is very efficient.

The useful patterns found in text mining process can be used to analyze text [8], and it also describes issue that identifying synonyms and polysemy through term-based approach is not possible. Assigning weight to specific pattern according to distribution of pattern is the main idea of this prototype model, and it improves the results of text mining process. A crime detection system is implemented using text mining tools, and relation discovery algorithm was designed to correlate the term with abbreviation [14]. A top-down and bottom-up approach of text mining process for web data is presented [15]. The text documents and their grouping are done by applying k-mean clustering methods, which is an unsupervised learning. The collected text documents can be grouped, the groups known as cluster by k-means clustering and other methods like agglomerative clustering. The similarity within text document collection can be calculated using TF-IDF algorithm, and this way relevant information is retrieved [1, 16].

In general, documents may be structured; semi-structured; or unstructured, but the majority of content is available in unstructured format [16]. To extract information from large collected of texts, documents are very complex and time-consuming task. A generic framework is discussed to mine concept visualizing phases of text improvement and knowledge distillation. The entity and its representation depend on specific use. The effectiveness of the extracted text data is improved by dividing text data into linear and nonlinear polynomial form by implementing support vector machine for classification. The experiments were performed to prove classification by suing multi-word features on the text.

## 3 Text Mining Techniques

Text mining(TM) and data mining are similar in many ways; the only difference lies in text mining uses unstructured text data. The unstructured data could come from variety of sources and available in form of HTML files, electronic documents,

e-mails, or texts on web pages. Text mining is an important process for organization or enterprises dealing with data in unstructured formats. The main techniques utilized in the process of text mining are information extraction, information retrieval, natural language processing (NLP), text summarization, and text categorization and clustering.

### **Information Extraction**

The collection of techniques that extracts meaning information in proper format or tabular format from unstructured text data is called information extraction [17]. The specification of attribute and relation can be given according to field from where text data are collected and this is done by domain experts [18]. The document collection content could be text, images, tables, and all kinds of plots; extracting relevant entities and features is the task of most IE systems. This way the corpus is extracted to be stored in database for further processing by text mining systems. Before extraction, detailed information about the desired entities should be known for better results [19]. The relation extraction is another task, which is to find semantics relation between entities [20, 21].

### **Information Retrieval**

The extraction of relevant patterns or information according to set of keywords or phrases provided by user is termed as Information Retrieval [22]. Both text and information retrieval is similar because the result is useful information according to the user need. Most IR system is used for retrieval purposes keep track of user's behavior and then returns relevant information on the basis of search keywords [19]. The search engines apply algorithms to find relevant documents or a page that fulfills user's current information need [1, 10]. The search engines like Google and Yahoo are extensively applying information retrieval system to retrieve content on basis of provided keyword by user.

### **Natural Language Processing**

Natural language processing (NLP) takes unstructured text data as input and automatically processes it to find hidden structure and patterns [10]. The NLP is the collection of software tools used to understand natural languages in written or verbal form. The analysis such as Named Entity Recognition (NER) is done to extracts synonyms for various entities in search. NER also helps in finding the relationships among entities and implemented using techniques like NLP [8]. NER identifies instances of all kind belonging to a specific object from a group of documents. The entities can find out the relationship to discover important concept in documents. However, this technique lacks complete dictionary list for all named entities used for identification [8, 12]. The complicated query-based algorithms can be used to get acceptable results [9].

### **Clustering**

Clustering is one of the text mining processes that group the text documents and known as unsupervised learning [16]. The similar terms are grouped to form clusters

extracted from text documents. The different clustering approaches are distance-based, hierarchical, density-based, and grid-based clustering method. The k-mean clustering method is example of distance-based approach [23]. In hierarchical clustering, a similarity function is used to calculate distance between text documents. To understand logical relationship and to identify role of a person in an organization can be handled using NER and co-referencing. Example is using name of a person at one place and using pronoun at another [23].

### **Text Summarization**

The process of producing short summary of original text documents is known as text summarization [24]. This process goes through steps like preprocessing and performing various text mining operations on the raw text to extract summaries [25]. The typical preprocessing steps are tokenization, stop word removal, and stemming methods that should be done before any other step. At the processing phase of summarization, first lexicons are generated. On the basis of the occurrence of a certain word or term in a text document, automatic summaries are created. Some other text mining methods were given to improve the accuracy and relevance of results [14]. Features like words, phrases, sentences, and paragraphs can be implemented to generate automatic summaries. The weighted heuristics method extracts features to summarize text documents [26].

## **4 Text Mining Applications**

Text mining applications offer opportunities in improving efficiency of all kinds of information retrieval tasks. Text mining processes unstructured text data to discover useful structure, hidden pattern, and sequences within the text. Text mining applications give power to organizations to find interesting pattern, knowledge, models, trends, future events, and rules from available unstructured data sources.

### **4.1 Life Science**

Healthcare and life science industries are producing huge amount of textual and numerical data like patient's record, diseases, medicines, symptoms, prescription, and treatments [24]. To filter out or retrieving an appropriate and relevant text for decision-making from a large text database is a difficult task [16, 27]. The health records collected by hospitals system contain complex diagrams and complicated medical terms making knowledge discovery very difficult [28]. Text mining tools in biomedical field are very helpful in retrieving useful information and relationship among various genes, treatment, diseases, and species. In medical field, text mining tools help in evaluating medical treatments and compare diseases, different symptoms, and treatments applied on people [16, 24, 29].

## ***4.2 Social Media***

The software packages for text mining are useful for analysis of social media data and monitoring. The analysis of content from blogs, e-mail, news, reviews, and twitter is possible using such software packages. These tools can easily analyze various posts on twitter, Facebook, and suggestions on social media network. It shows users interests belonging to specific age group or their views about the same post [30, 31]. Twitter, a social media platform, generates texts on various topics, which are useful for predicting future events and trends [32, 33]. Another very important application is predicting sentiments of people toward a product or a person [32].

## ***4.3 Business Intelligence***

Text mining is very important in business analytics and helps enterprises to analyze how their competitors are performing. Also, it analyses customers purchase patterns and thus improving decision-making. It gives an overall picture of the business and also gives information to increase customer satisfaction and to have a competitive advantage over competition enterprises. IBM text analytics and Rapid miner are text analytics tools primarily used in business intelligence [32]. IBM text analytics helps in taking decisions that give response related to performance of the organization. Telecommunication sector, commerce applications, business organization, logistics, and customer relationship management system utilize such methods to improve their process and profit [33].

## ***4.4 Education and Research***

The various techniques and tools popularized by text mining are used to do analysis of latest trends in education sector. These tools help in understanding student's need and employment ratio [26]. The text mining techniques are being used in classifying research papers and in retrieval of interesting content for the user. The clustering methods like k-means can be used to identify relevant attributes according to data analysis needs. The performance of students in examination for different subjects can be summarized and assessed. The selection of various subjects by people and affects of various parameters can also be analyzed [14, 18, 34].

#### **4.5 Libraries Online**

The techniques of text mining can be applied to explore patterns and trends in online journals, proceedings, and articles [27]. These online resources help in so many ways to researcher or any other user in need for information. Such libraries are large repositories of information; also, digital libraries are really significant to available collection retrieval. It also provides organizing ability making it possible to store all kinds of data. The Green-stone digital library provides a multilingual interface for extracting documents that handles formats like Microsoft word, pdf, postscript, HTML, scripting languages, and e-mail messages [14]. This system provides support in document extraction available in audio visual forms, image, and text data.

#### **4.6 Financial Systems**

Financial information mainly lies in financial statements [4, 17], but it should be noted that financial information can be textual and numerals. There are many data sources (footnotes, sustainability reports, executive letters, etc.) that are different from the financial statements that provide useful valuable information for decision makers. In recent times, numerical data structure of the financial statements started to become insufficient in supporting business decisions of the stakeholders. Thereby, studies on text mining of financial information have begun to raise with the increase in the text data. Understanding the text data contained in the financial reports produced by enterprises is very important for financial information research.

#### **4.7 Electronic Discovery**

The electronic discovery [35] and its effectiveness in information retrieval have been linked to judicial rulings and practitioner controversy. The E-discovery tasks and its growing demand are pushing information retrieval system to their limits. The growth of “Electronically Stored Information” (ESI) is making it difficult for the government agencies to store and retrieve as they are required to retain all the legal documents [22]. The constant growth, size, and complexity of electronically information are subjected to routine capture in litigation and put serious challenges on Information Retrieval (IR) researchers. There is a fundamental question as to how best to model the real world should be answered by them [36].

#### 4.8 *Online Advertizing*

Study participants of advertizing agencies often bid for advertizing space next to searches for chosen keywords [36] as firms do it in their advertising campaigns. The Amazon, Facebook, Google Adwords, Bing Ads, and other such online organization need it. These platforms keep a list of keywords or suggested keywords for such clients who are spending money by mining information from sources like click stream data and web logs from various website by users. The searched keywords can be extracted from search engines query logs and most visited web pages. Text mining helps in online advertizing by automatically selecting most appropriate ads for a specific web page. Ads can also be selected on the basis of query submitted on page by user [25].

### 5 Text Mining Issues

Issues taking place during text mining process directly affects the effectiveness of decision-making and sometimes incorrect results. Many issues or complexities can happen at text mining intermediate phases. For example, in preprocessing phase rules are defined to make TM process efficient. Before doing any processing, all the contents from various document are converted into intermediate form then only other steps of text mining is performed. Sometimes data are modified during preprocessing phase, and so it loses its importance [28]. Text refinement dependency in case of multilingual documents is also a major hurdle in mining of such documents as only few tools are available which processes multiple languages [37]. Another issue is text representation; a document should be represented by single word or multiple words, by a phrase or by a sentence.

There exist many techniques, which support multilingual text like English, Japanese, German, Hindi, and many others. Numerous text documents and their processing are difficult because tools are not available to support them. Current techniques and software have limitations when it comes to support multilingual documents. Domain knowledge and integrating it with text mining is also important considering impact on final outcomes. The experts from their respective field have to work collaboratively to extract accurate results [23, 28].

The synonyms, antonyms, and polysemy usage in text documents make it difficult for text analysis tools to produce result correctly. When document is very large and collection is from a diverse field, it is very difficult to categorize the document. The previous work has shown that recognizing and classifying named entities in texts require knowledge on the domain entities. List entities are used to tag text entities, with the relevant semantic information; however, exact character strings are often not reliable enough for precise entity identification [22]. As elaborated in [35], the ambiguity is still the major “world problem” in text mining applications.

The acronyms can have different meaning at various places is a common problem in text mining [17]. The granularity concepts change the context of text depending on domain knowledge. The plug-ins is embedded in text mining tools as per the specific field standards and rules. Such plug-in can be developed by collecting proper knowledge about specific domain [32, 37]. The processing of natural language documents faces lots of complexities in text refinement methods and in identification of named entities. Same spelling words can have different meaning, for example, “set” and “set.” The first set means “to arrange” while other mean “collection.” Many text mining tools will consider both as similar while one is verb, the other is noun. Rules of grammar according to context of use and nature of text are also major issue and still an open problem [25, 37, 38]. Users working on retrieval system face difficulty in recalling a high percentage of relevant keywords but can easily recognize important keywords when given to them as options [39]. A randomly ordered keywords list that is provided to users helps them in retrieval of relevant documents.

Processing of medical records is difficult because the terms they contain are ambiguous many times. Word Sense Disambiguation is the technique of identifying the actual and correct meaning of an ambiguous word according to the given context. Many applications are being used by organization for handling such ambiguities; such systems are popularly known as Word Sense Disambiguation (WSD) systems [38]. The term “bank” could refer to the financial institution or the river “bank,” and meaning is totally dependent on the context of use. Another important issue is occurrence of unknown word that might appear in text documents like some local slang. Many text analytics tools like text categorization and sentiment analysis need annotated data, but there is lack of such data [25].

Recognizing entities and their instances is another issue; for example, any single entity can have numerous terms like USA, US, America, or North America represents same entity. In many places in text, multi-words names are given to group of successive words to identify boundaries and possible overlapping. This is done using classifications technique [40]. Recognizing named entities is an issues too as same object or entities can have many names. Most NER systems have achieved 75 percent to 85 percent relevance level. A technique known as co-referencing is used for NLP [12, 13]. The inherent complexities in most natural languages extracted from disparate sources may not have identical words for same person, for example, Mahatma Gandhi, Gandhi, and Mohan Das Karamchand Gandhi.

## 6 Conclusion

The large-scale text databases need to be explored and analyzed to fulfill user’s current information need helping them in decision-making. This new found information can further be used in decision-making and predictive modeling. The selection and use of appropriate technique fulfilling current requirement makes the process of mining large-scale text corpus efficient. The application area of text mining is very diverse including life science, research, online libraries, online advertizing, financial systems,

and business intelligence. Primary issues in processing of large-scale text data are multilingual text, presence of ambiguities in natural languages and integration of domain knowledge. The ambiguities lie in form of synonyms, polysemy, acronyms, and word sense. These issues need to be handled during text mining process failing that could give rise to incorrect results. In course of my continuous research, our primary focus will be on designing algorithms to resolve issues presented in this paper.

## References

1. Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut KA (2017) brief survey of text mining: classification, clustering and extraction techniques. arXiv preprint [arXiv:1707.02919](https://arxiv.org/abs/1707.02919)
2. Padhy N, Mishra D, Panigrahi R et al (2012) The survey of data mining applications and feature scope. arXiv preprint [arXiv:1211.5723](https://arxiv.org/abs/1211.5723)
3. Fan W, Wallace L, Rich S, Zhang Z (2006) Tapping the power of text mining. Commun ACM 49(9):76–82
4. Rajendra R, Saransh V (2013) A novel modified apriori approach for web document clustering. Int J Comput Appl 159–171
5. Weiss SM, Indurkhya N, Zhang T, Damerau F (2010) Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media
6. Gupta V, Lehal GS (2009) A survey of text mining technique and applications. J Emerg Technol web Intell
7. Liao S-H, Chu P-H, Hsiao P-Y (2012) Data mining techniques and applications—a decade review from 2000 to 2011. Expert Syst Appl 39(12):11303–11311
8. Welbers K, Van Atteveldt W, Benoit K (2017) Text Analysis in R. Commun Methods Meas 11(4):245–265
9. Cohen AM, Hersh WR (2005) A survey of current work in biomedical text mining. Brief Bioinform 6(1):57–71
10. Manning CD, Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT Press
11. Li J, Sun A, Han J, Li C (2018) A survey on deep learning for named entity recognition. arXiv preprint [arXiv:1812.09449](https://arxiv.org/abs/1812.09449)
12. Henriksson A, Moen H, Skeppstedt M, Daudaravičius V, Duneld M (2014) Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semant 5(1):1
13. Kaur H, Chauhan R, Alam MA, Aljunid S (2012) SpaGRID: a spatial grid framework for high dimensional databases, pp 690–691
14. Chen CP, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf Sci 275:314–347
15. Heidari M, Felden C (2015) Financial footnote analysis: developing a text mining approach. In: Proceedings of international conference on data mining (DMIN), pp 10–16
16. Murtagh F (1983) A survey of recent advances in hierarchical clustering algorithms. Comput J 26(4):354–359
17. Nedellec C, Nazarenko A (2005) Ontologies and information extraction: a necessary symbiosis. In: Buitelaar P, Comiano P, Magnin B (eds) Ontology learning from text: methods, evaluation and applications. IOS Press Publication
18. Chauhan R, Jangade R, Rekapally R (2018) Classification model for prediction of heart disease. In: Soft computing: theories and applications, pp 707–714
19. Athenikos SJ, Han H (2010) Biomedical question answering: a survey. Comput Methods Programs Biomed 99(1):1–24

20. Kaur H, Chauhan R, Afshar Alam M (2010) Spatial clustering algorithm using R-tree. *J Comput* 3(2):85–90
21. Chan YS, Roth D (2010) Exploiting background knowledge for relation extraction. In: Proceedings of the 23rd international conference on computational linguistics . Association for Computational Linguistics, pp 152–160
22. Oard DW, Baron JR, Hedin B (2010) Evaluation of information retrieval for E-discovery. *Artif Intell Law* 18:347
23. Kaur H, Chauhan R, Aljunid SM (2012) Data mining cluster analysis on the influence of health factors in Casemix data. 12(suppl 1):2–3
24. Chauhan R, Kumar N, Rekapally R (2019) Predictive data analytics technique for optimization of medical databases. In: Proceedings of SoCTA, pp 433–441
25. Ittoo A, Nguyen LM, van den Bosch A (2016) Text analytics in industry: challenges, desiderata and trends. *Comput Ind* 78:96–107
26. Al-Hashemi R (2010) Text summarization extraction system (tses) using extracted keywords. *Int Arab J e-Technol* 1(4):164–168
27. Witten IH, Don KJ, Dewsnp M, Tablan V (2004) Text mining in a digital library. *Int J Digit Libr* 4(1):56–59
28. Henriksson A, Zhao J, Dalianis H, Boström H (2016) Ensembles of randomized trees using diverse distributed representations of clinical events. *BMC Med Inform Decis Mak* 16(2):69
29. Alonso I, Contreras D (2016) Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach. *Expert Syst Appl* 44:386–399
30. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* 3(02):185–205
31. Zhao Y (2013) Analysing twitter data with text mining and social network analysis. In: Proceedings of the 11th Australasian data mining and analytics conference
32. Dörre J, Gerstl P, Seiffert R (1999) Text mining: finding nuggets in mountains of textual data. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 398–401
33. Sharda R, Henry M (2009) Information extraction from interviews to obtain tacit knowledge: a text mining application. In: AMCIS 2009 proceedings, p 283
34. Ayesha S, Mustafa T, Sattar AR, Khan MI (2010) Data mining model for higher education system. *Eur J Sci Res* 43(1):24–29
35. Sanderson M, Zobel J (2009) Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, pp 162–169
36. Antoun C, Zhang C, Conrad FG, Schober MF, Comparisons of online recruitment strategies for convenience samples: craigslist, Google AdWords, Facebook, and Amazon mechanical turk 28(3):231–246
37. Türegü N (2018) Text mining in financial information. In: Current analysis on economics & finance, pp 18–26
38. McInnes BT, Stevenson M (2014) Determining the difficulty of word sense disambiguation. *J Biomed Inform* 47:83–90
39. King G, Lam P, Roberts M (2014) Computer-assisted keyword and document set discovery from unstructured text 456
40. Wen Z, Yoshida T, Tang X (2007) A study with multi-word feature with text classification. In: Proceedings of the 51st annual meeting of the ISSS-2007, Tokyo, Japan, vol 51, p 45

# Vehicle Number Extraction Using Open Source Tools



Chetan Pandey, Amit Juyal, and Ankur Dumka

**Abstract** Although intelligent monitoring and recording system (IMRS) is used in many fields such as aviation traffic control, transportation, real estate, medical science and more. One of these is vehicle traffic on roads. This paper covers the techniques of extraction of vehicle number by using open source tools which will be used in many fields such as automated parking system, toll tax system, supervising road traffic on highways, searching stolen cars, and more. Time and accuracy are two challenges in performing the above in a real-life scenario. The proposed vehicle number extraction works on OpenCV Library to convert the image to Mat, and after that, the discussed methodology is applied to detect the vehicle number. This paper introduces a framework which includes preprocessing of vehicle image and applies non-maxima suppression with edge detection and segmentation, and then, OCR has been applied to convert the numbers to digital form and store in HBase database along with an image of the car. It has been observed that the proposed methodology of this paper runs 27% faster than automatic number plate recognition (ANPR). Precision and recall are also observed to be better.

**Keywords** OpenCV · Image processing · OCR · Tess4J · NMS

---

C. Pandey (✉) · A. Juyal  
Graphic Era Hill University, Dehradun, India  
e-mail: [schetanpandey@gmail.com](mailto:schetanpandey@gmail.com)

A. Juyal  
e-mail: [amitjuyal26@gmail.com](mailto:amitjuyal26@gmail.com)

A. Dumka  
Graphic Era Deemed to be University, Dehradun, India  
e-mail: [ankurdumka2@gmail.com](mailto:ankurdumka2@gmail.com)

## 1 Introduction

Detection of vehicle number plays a very important role in IMRS [1, 2]. Intelligent monitoring covers supervision of the target system like road traffic, aviation traffic, the stock market, some parking areas, and more. This will be achieved by using cameras, sensors, e.g., proximity sensor or any suitable scientific tools. Recording system implies a collection of images, videos, or the outcomes of tools used for monitoring. This paper focuses on the vehicle traffic and will majorly focus on the extraction of vehicle number from the vehicle and store them in some text file format.

In India, basically there are two types of license plates: One is black characters in a white plate (private vehicles like bike, and car), and another is the black characters in a yellow plate (generally public transport like taxi). This becomes challenging for capturing images or making videos of these vehicles. White does not make much trouble, but yellow color may since when some light fall over it, and it becomes too much bright making the number to appear dull in the image. Also, style of writing numbers over license plate is different; i.e., some plates have the number in one line and some have two lines in the license plate. However, the recognition task becomes challenging if the color of vehicle number is very similar to the background since some uses multicolor in vehicle numbers.

In this paper, capturing of vehicle will be done by using good quality of digital cameras such as DSLR-based. However, due to human error and natural factors, many images are not clear; i.e., it is difficult to extract vehicle number from them. Also, the vehicle owner may use a different style of writing vehicle number and also different countries have different syntax of vehicle number plate—all these factors create disturbance and noise while collecting vehicle number. This paper will discuss a user-oriented approach to remove noise and the other disturbing objects from the image after applying digital image processing methods [1]. The basic techniques are gray-scale image, edge detection, dilation, and erosion. If required de-blurring algorithms will be used to correct blurred and defocused vehicle images. After that, character segmentation [3] is used, and finally, license number is extracted from this license plate image using some template matching technology like optical character recognition (OCR) [4, 5].

## 2 Related Work

Detection of vehicle number from vehicle image is although too old research area but still due to its importance and value in another area of research, like stolen car detection, automated parking system, and more, it still an interesting and evergreen area of research. There are many methods to extract license's character like some paper shows the use of neural network [6], support vector machine [6], OCR [7], Matlab [2], Hough transform, soft computing, and more. Like in the paper [6] published in the year 2000, here authors for image segmentation used neural network and

for character recognition, they focused on support vector machine. Another paper [8] published in the year 2004 makes use of soft computing techniques for vehicle number detection.

Focusing on the pre-processing [9], in the year 2014, a paper [2] is published in which authors suggest the techniques for extraction of license number from the vehicle number plate. This paper shows the use of edge detection and morphological operations, especially dilation as a suggested method. Finally, optical character recognition is used with template matching algorithm for finding license number.

In 2014, another paper [3] preferred free and open source tools like OpenCV over license-based tools like MatLab. In this paper, authors uses Python programming language along with the Open Computer Vision Library to form automatic license plate recognition system. Through experiments, the paper shows that how OpenCV [10, 11]. Library is useful and far better than Matlab for processing images as well as videos. Basically, here authors want to show that open source tools are enough helpful for scientific computing domains.

Many research papers show some automatic technology of detecting numbers from vehicle images. One of the papers [1] is published in the year 2014. Here, authors had designed an efficient, effective, and easy to use vehicle identification system which is fully automatic. As the author mentioned, ANPR as an image processing technology uses color number plate as inputs and the resulted output is the license number of that vehicle. It automatically senses either the front view or back view of the vehicle. Basically, the whole system is divided into four main steps by the author, i.e., image acquisition, plate localization, character segmentation, and recognition [3]. Matlab 2010a is used here to implement and simulate this proposed system.

There are some research papers focusing on some different technology like optical character recognition (OCR) to not only detecting license number but also storing them in some text format at the same time. One of such paper [10] is published in the year 2015. In this paper, authors presented a methodology of extracting the number from the captured vehicle image by using Java OCR Libraries. Here, they suggested the ASPIRE OCR-java library [10] to extract all information from the image into editable text format. Though, according to authors, the experiments showing positive result but still there are some issues like stains, blurred regions, and differences in font styles and sizes need to be resolved properly.

In 2017, there is another paper [9] focuses on the use of OCR technology along with template matching algorithm [1]. Here, authors divide their work into two parts. The first part is number plate area extraction, and the second part is character identification. For first part, morphological operations are used. After the extraction of the number plate, character segmentation is used after the first part which is based on the histogram approach. After this, template-matching method is used for character identification. The described methodology in this paper is for the image. However, video-based surveillance system needs to be implemented further.

### 3 Proposed Framework

The most important aspect of this work is clear and accurate vehicle image as much as possible. If still there are some issues remain with the image, then that will be corrected by using basic digital image processing methods [9] which will be discussed further in this paper. So for good vehicle image, use good resolution digital camera (suggesting minimum resolution will be  $24 * 42$  pixels). Always check that the number plate must be completely visible which clicking image.

This paper will further discussed following steps for vehicle number extraction:

- Convert vehicle image to gray-scale image
- Edge detection in resulted image
- Applying dilation operation
- Character segmentation
- Apply optical character recognition (OCR) (Fig. 1)

#### 1. Convert vehicle image to gray-scale image

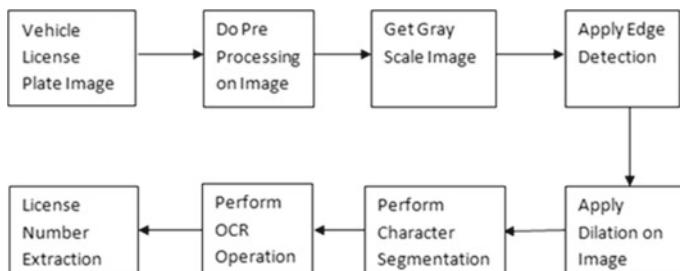
The basic aim of this step is to enhance the clarity of the vehicle image. Generally, images are in RGB format. In RGB [1], each pixel has a composition of three color, i.e., red, green, and blue. To convert an RGB image to grayscale [8], we need to convert all RGB values into grayscale values for each pixel. Technically, we take the sum of red, green, and blue after multiplying them with their respective grayscale values, i.e.

$$\text{Red\_Value} * 0.299 + \text{Green\_Value} * 0.587 + \text{Blue\_Value} * 0.114$$

Take care that vehicle number plate must be clearly visible and must have no blurriness, and if somehow it has then first deblurred the image, then convert it to grayscale.

#### 2. Edge detection in resulted image

In general term, edge detection is finding the boundaries of an object within the image. It works on the basis of the discontinuities in the image brightness by forming an edge



**Fig. 1** Methodology of vehicle number extraction

between two pixels (point). Since we are using OpenCV Java Library [11], it has many predefined Edge Detection Operators like Sobel, Canny, Prewitt, and more. Through experimenting all these operators, this paper suggests the Canny edge detection operator [2, 3]. In OpenCV, it is defined as:

```
public static void Canny (Mat image, Mat edges, double threshold1, double
threshold2, int apertureSize, boolean L2gradient)
```

Canny generally uses gray-scale image as the input image. ‘Threshold 1’ and ‘threshold 2’ are used for linking edge between two pixels (point). Both values should be moderate since if it is too low, then it may detect edges of unnecessary objects and taking too high value may not detect edges of important objects within the image. ‘ApertureSize’ is used for Sobel operator. ‘L2gradient’ is used to set image gradient magnitude, i.e., either TRUE or FALSE. Considering the intensity of an image, a change in the direction is often referred as image gradient. Gaussian filter (1) is applied to the image to smoothen it; then, intensity gradients are found (2), and then, applying non-maxima suppression, spurious edges are removed. Then, potential edges are determined to apply a double threshold; finally, strong and connected edges are marked by suppressing weaker unconnected edges.

$$\begin{aligned} Gf_{xy} &= \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x - (k+1))^2 + (y - (k+1))^2}{2\sigma^2}\right); \quad 1 \leq x, y \leq (2k+1) \\ Gf_x &= \frac{\delta}{\delta x}(f * G) = f * G_x \\ Gf_y &= \frac{\delta}{\delta y}(f * G) = f * G_y \end{aligned} \quad (1)$$

$$G\text{magnitude}(i, j) = \sqrt{Gf_x^2 + Gf_y^2} \quad (2)$$

#### Non-maxima Suppression:

- Local maxima [12, 13] of the gradient magnitude is first calculated.
- Thinning of broad ridges is carried out to ensure that only points to maximum change remain.
- Smaller values of a ridge along the direction of the gradient are suppressed.

#### Algorithm for applying edge detection in an image

Step1. begin

Step2. for image(pixel)

Step3. if  $G\text{magnitude}(x, y) < G\text{magnitude}(x_1, y_1)$

or  $G\text{magnitude}(x, y) < G\text{magnitude}(x_2, y_2)$

then  $In(x, y) = 0$

else

$In(x, y) = G\text{magnitude}(x, y)$

Step5. end if

Step4. end for

Step6. ends

### 3. Applying dilation operation

In general, dilation refers to focus on the edges of all the objects of an image by adding more pixels to it. The basic rule for dilation [4] is that the value of the adding pixel must be more than the values of its neighboring pixels. Dilation uses the value of structuring element which is generally taken as a rectangle. The value defined here is scanned over the boundaries of objects; for each element, we find the maximum pixel value among the neighboring pixels and replace the pixel value of that element with the maximum value. Simply, it causes bright region in the image to grow more resulting in more clarity in the foreground of objects. In OpenCV, it is defined as

```
public static void dilate(Mat src, Mat dst, Mat kernel) (3)
```

‘src’ contains source file, ‘dst’ contains output file, and ‘kernel’ is a structuring element that is used while dilation is applied. The structuring element is achieved by using ‘getStructuringElement’ method which found under ‘Imgproc’ class. This function is used to dilate the source image, which is used to pick the maximum pixel value among the neighboring pixels.

### 4. Character Segmentation

Character segmentation [3, 9] is used to remove unnecessary stuff from the resulted license plate image. This makes easy to pick license number from the image and to store it later. However special care must be taken while segmentation since we need only that region which contains license number and leaving the other stuff like bolt’s over the license plate, branding logos, and more. Keep in mind that segmentation is very important step among rest since the failure of it can lead us to an ambiguous result and we will not successfully extract the license number from the license plate.

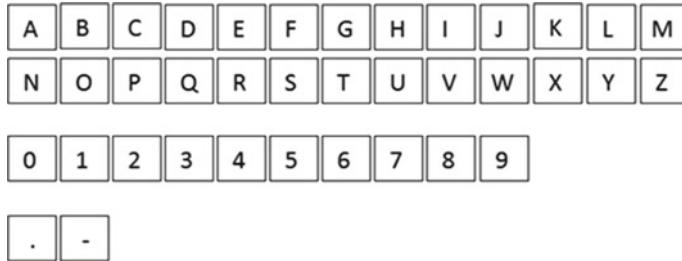
Within a word, each character has a horizontal space called ‘interword space’, while the horizontal area between word is called as ‘Interletter space’. Each alphabet and numbers are segmented into the processed image and saved separately. This is used in the OCR techniques like Template Matching Algorithm or Feature Detection.

### 5. Optical Character Recognition (OCR)

OCR [4, 14] extracts all alphabets, numbers, and some symbols basically ‘.’ and ‘-’ from the resulted Image into a text format. OCR is very useful in Passport Offices, Banks (especially for passbook printing), Shopping Receipts, and more. In this process, characters are selected from the image after filtering and then matched with predefined characters using OCR techniques [1] like Template Matching Algorithm. Note that the predefined characters must be in image form containing alphabets (capital letter and small letter), numbers (0–9), and special symbol (space, hyphen).

#### 1. OCR Working

In license plate, the font style varies resulting in a different style of writing license number. If it is simple for example alphabets, then it is manageable but the challenging



**Fig. 2** Predefined characters for template matching algorithm

part is to recognize the simple as well as Italic characters or stylish/fancy characters. Considering this problem, there are two OCR techniques: One is pattern recognition, and another is feature detection.

In pattern recognition [9, 10], the segmented image compares with the stored characters' image, and if they matched, the result will be stored in text format. This is done by using Template Matching Algorithm [1]. But for such process, the written characters should of nearly same font style and size. Such techniques are highly used in bank checks, demand draft, and more. Every letter is of exactly same width, and the strokes were designed in such a way that each letter would be easily identified by all the others (Fig. 2).

Feature detection, also called intelligent character recognition (ICR), is much user oriented for spotting characters. It uses set of rules according to each alphabet, each number, and each special character. For example, a round shape is 'O'/'o' alphabet, if it is round but the upper portion is not connecting this means 'U'/'u' alphabet and more rules. So using such features as a rule makes it easy to recognize all characters no matter what is the font style and size. Thus, in this technique, instead of matching the complete pattern of characters, only individual character's features are detected which makes this technique fast and user-friendly.

Feature detection approach by using Tess4J library [15] contains a very effective method to apply OCR [14] in vehicle image with the help of Tesseract API [15]. The recognition process in this API is a two-pass process. In first pass, all characters are tried to recognize with the help of an adaptive classifier. However, due to a different type of font style, some characters are not well recognized in the first pass. So in second pass, those characters are recognized again by checking alternative hypothesis.

## 2. Algorithm for applying OCR in an Image

Step1. begin

Step2. for image(jpg, png)  $\leftarrow$  FileNameExtension

loc. of image  $\leftarrow$  JFileChooser

JFileChooser  $\leftarrow$  apply FileNameExtension

select image I  $\leftarrow$  JFileChooser.getSelectedFile

apply Tesseract.doOCR(I)

Step3. end for  
 Step4. end

## 4 Observation of Proposed Methodology

Following are the steps of this proposed framework for the vehicle number extraction:

Step1.  $I \leftarrow$  read image  
 Step2.  $M \leftarrow$  convert to matrix  
 Step3.  $I_{gray} \leftarrow$   $rgb(M(\text{Red\_Value}*0.299+\text{Green\_Value}*0.587+\text{Blue\_Value}*0.114))$   
 Step4.  $\text{CannyedgeImage} \leftarrow$  cannyEdge( $I_{gray}$ ) // (using (1) and (2))  
 Step5.  $\text{dilatedImage} \leftarrow$  dilate( $\text{CannyedgeImage}$ ) // (using (3))  
 Step6.  $\text{lineFreeImage} \leftarrow$  removelines( $\text{dilatedImage}$ )  
 Step7.  $\text{SegmentedPlate} \leftarrow$  segment( $\text{lineFreeImage}$ )  
 Step8.  $\text{Platenumber} \leftarrow$  OCR( $\text{SegmentedPlate}$ ) // tess4j is used for OCR

For observation, firstly we need to set up computer's environment, i.e.,

- Install Java (jdk1.8.0\_144)
- Install Eclipse (Oxygen Release (4.7.0))
- Install OpenCV (opencv-3.3.0-vc14)
- Install Tess4J (Tess4J-3.4.2-src)

Within this environment number of vehicle license plate, images are tested successfully and properly extracted the vehicle's license number into a text document.

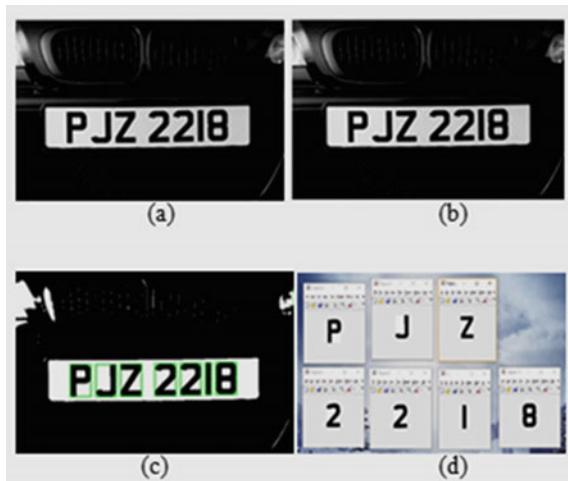
Vehicle number extraction can be achieved in a number of ways, and a number of java functions are available like for binary image or for indexed image. In OpenCV, there are predefined functions related to digital image processing methods such as grayscale conversion, edge detection, and more. There is 'Mat' [7, 10] class defined in OpenCV which store images as indexed image, i.e., in matrix form. Also, many researchers suggested that digital processing gives a better result if an image is in matrix form. And in OpenCV, there are lots of functions associated with 'Mat' class. This paper suggests using 'Mat' class, and its functions as described above in the methodology part.

While following the suggested methodology, it observed that it is very simple and user-friendly to use OpenCV [3, 7] library along with Java. This also does not require too much of coding. For example, for RGB to grayscale conversion, 'Mat' class has 'Imgproc' class which has 'cvtColor' function for this, i.e.,

```
Imgproc.cvtColor (Mat_Source_File, Mat_Object_to_store,
Imgproc.COLOR_RGB2GRAY);
```

For edge detection, Sobel, Prewitt, Canny, and more operators are tested, and out of which, Canny edge detector gives a better result. However, which edge detection is better is completely dependent on the target object. Since basically, it is finding object boundaries, and here, license plate is our target object and through our experiments,

**Fig. 3** **a** Grayscale image,  
**b** canny edge detection,  
**c** dilated image, and  
**d** character segmentation



Canny gives much better result compared to others. One thing which makes user-friendly than others is that it allows the user to adjust ‘threshold’ values accordingly and can further adjust the result by using different values and setting the ‘gradient’ value. Means a lot of option for getting the desired image.

For OCR, there are two approaches: One is Template Matching Algorithm and another is Tess4J library [15]. Note that this is the final and most important step in getting the license number of a vehicle. Rest of the above-defined steps are basically used to get clear, accurate, and focused license plate image. But the later step is strictly for getting all characters presented in the input image. If somehow pattern recognition, i.e., template matching fails, then too we have intelligent character recognition (ICR), i.e., Tesseract API [15] in OCR, which used recognition related rules for each character. So not only it is helpful in defining a character’s characteristics but also extracts that and convert that character into ASCII code which is then easily stored in text format. The output of the suggested steps are (Fig. 3).

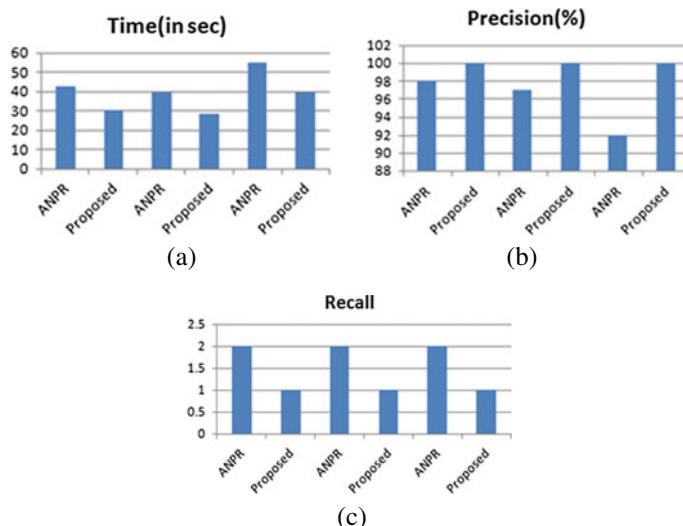
## 5 Result

The proposed methodology is giving a satisfactory result with a different type of images along with different image size. For comparison, the paper opted the ANPR which is an automatic vehicle number identification system. Automatic number plate recognition (ANPR) is a technique of automatically processing the image for extraction of vehicle number. It uses the color image of the vehicle, and then, it automatically senses the license plate in the input image and shows the number printed on the vehicle image as an output.

The proposed vehicle number extraction methodology is compared with the ANPR, and it is observed that the prior giving a better result than the later. Robustness of the proposed system is checked on the basis of two well-known classification parameters, i.e., precision and recall for different images and are calculated using (Fig. 4 and Table 1).

$$\text{Precision} = \frac{\text{no. of relevant result}}{\text{total no. of retrieved results}}$$

$$\text{Recall} = \frac{\text{no. of relevant results}}{\text{total no. of expected results}}$$



**Fig. 4** Graphs showing comparison between ANPR and proposed methodology in this paper on the basis of time, precision, and recall, respectively

**Table 1** Results obtained of time, precision, and recall

Technique	Image size	Type	Time (in sec)	Precision (%)	Recall
ANPR	128 × 128	jpg	42.54	98	2
Proposed	128 × 128	jpg	30.6288	100	1
ANPR	128 × 128	bmp	39.20	97	2
Proposed	128 × 128	bmp	28.2240	100	1
ANPR	256 × 256	jpg	55.054	92	2
Proposed	256 × 256	jpg	39.6388	100	1

## 6 Conclusion

The discussed methodology of detecting license number of vehicles and storing them in the text format is precisely tested into Windows environment by implementing it in Eclipse using Java as programming language along with OpenCV Library. All experiments are done using real-time vehicle images. All suggested steps are attentively implemented over the vehicle images and can be implemented on alive video recording of the vehicle also. Whether the image is taken from the front side or back side or even if image quality is poor, it will still be able to extract the number from the image on applying the above discussed methodology. OpenCV Library and OCR API (Tesseract) with filtering the output using simple text mining steps make it possible to achieve good results. Apart from the discussed functions, user can use another function also if it is needed but must make sure to apply OCR as a final step. Once the license numbers are stored in text format, we can use these data in many applications like vehicle parking access control, automated toll tax calculator, detecting stolen vehicles, and more.

## References

1. Gaikwad DY, Borole PB (2014) A review paper on automatic number plate recognition (ANPR) system. *IJIRAE* 1:88–92
2. Lakshmi Priya V, Perumal K (2014) Detecting the car number plate using segmentation. *IJECS* 3:8823–8829
3. Jain P, Chopra N, Gupta V (2014) Automatic license plate recognition using OpenCV. *IJCCTR* 3:756–761
4. Qadri MT, Asif M (2009) Automatic number plate recognition system for vehicle identification using optical character recognition. In: ICETC. IEEE, Singapore, pp 335–338
5. Prabhakar P, Anupama P, Resmi SR (2014) Automatic vehicle number plate detection and recognition. In: ICCICCT. IEEE, Kanyakumari, India, pp 185–190
6. Kim KK, Kim KI, Kim JB, Kim HJ (2000) Learning-based approach for license plate recognition. *NNSP X*, vol 2. IEEE, Sydney, pp 614–623
7. Kaili C, Meiling W (2014) Image stitching algorithm research based on OpenCV. Chinese Control Conference, IEEE, Nanjing, China
8. Chang S-L, Chen L-S, Chung Y-C, Chen S-W (2004) Automatic license plate recognition. *TITS*, vol 5. IEEE Transactions
9. Suryanarayana PV, Mitra SK, Banerjee A, Roy AK (2005) A morphology based approach for car license plate extraction. In: India Conference, Indicon, IEEE, Chennai, India, pp 24–27
10. Madhu Babu D, Manvitha K, Narendra MS, Swathi A, Praveen Varma K (2015) Vehicle tracking using number plate recognition system. *IJCSIT* 6(2):1473–1476
11. Xia Y, Chen J, Lu X, Wang C, Xu C (2015) Big traffic data processing framework for intelligent monitoring and recording systems. *ScienceDirect* 181:139–146
12. Zhang L, Sun Y, Chen F (2015) An improved edge detection algorithm based on fuzzy theory. *FSKD*, IEEE, Zhangjiajie, China, pp 380–384

13. Non Maximum Suppression, <https://www.pyimagesearch.com/2014/11/17/non-maximum-suppression-object-detection-python/>
14. Optical Character Recognition and Detection of text in an image, <https://in.mathworks.com/help/vision/examples/automatically-detect-and-recognize-text-in-natural-images.html>
15. Tess4J and Tesseract, <http://tess4j.sourceforge.net/docs/docs-3.4/>

# Classification of Energy Efficiency in Mobile Cloud Computing



Shubham Pal and Ankur Dumka

**Abstract** Mobile cloud computing (MCC) is a methodology, which is developed due to the inability of mobile devices to process large amount of data and utilize less amount of energy as such the computers that can process the large amount data as compared to mobile devices. So in order overcome this problem, MCC came into existence which is used to increase the computation power and utilize energy of mobile devices that is required to process large data; to overcome this issue, there are several techniques that we discuss in this paper and their proposed solution to enhance the computation ability of mobile devices by using less energy. Techniques involve in taking off the data from mobile devices to the cloud server and perform the computation in cloud server, and when the computation of data is completed, then send back that particular data to the mobile devices. Thus, this paper studies about how to reduce the energy consumption of mobile devices by using certain parameters such as bandwidth and execution time.

**Keywords** MCC · Storage · Energy efficiency · Battery lifetime

## 1 Introduction

Mobile cloud computing (MCC) consists of mobile computing and cloud server computing. MCC is a collection of mobile devices, cloud, and wireless network like Internet that provides rich computation resources. Actually, the aim of MCC is to reduce the workload on mobile devices so that to operate large amount of data by using less amount of energy [1]. In mobile devices, because of its small physical size and energy constraint for the application, we are not performing computation of application on mobile devices due to its high energy usage. The capacity of storage

---

S. Pal (✉) · A. Dumka  
Department of Computer Science and Engineering,  
Graphic Era Deemed to be University, Dehradun, India  
e-mail: [spal5776@gmail.com](mailto:spal5776@gmail.com)

A. Dumka  
e-mail: [ankurdumka2@gmail.com](mailto:ankurdumka2@gmail.com)

of the mobile devices still falls behind the desktop. So in order to resolve this issue, MCC came into picture. In MCC, the computation of data is to be performed on cloud and desired output is sent to the mobile devices. In such a way, we can enhance the computation ability of mobile devices and improve its energy efficiency. Mobile devices have low battery life and low processing capability to improve its battery life and processing capability.

The problem is that how we can integrate the **storage service** in MCC and how we can save the **energy** of the devices when we interact with the cloud. In order to preserve the energy of mobile devices, we have to perform offloading but offloading is not always energy saving as compared to local computation. MCC can impact its capability to make mobile devices as powerful as desktop by offloading the data to the cloud, but this may not be always feasible due to the **limited battery lifetime** of mobile devices because offloading may result high power utilization due to communication overhead which favor local computation [2].

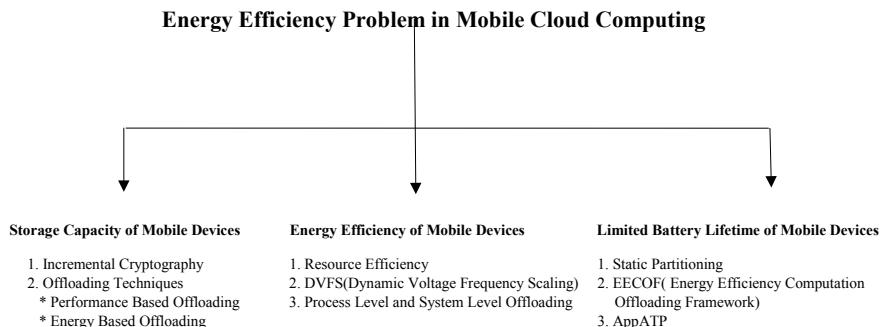
## 2 Review

Energy efficiency problem in MCC can be categorized into following sub-parts as shown in Fig. 1.

### Storage Capacity of Mobile Devices

For problem of storage capacity in MCC, there are many solutions provided by many researchers for this issue, of which we had discussed two as incremental cryptography and performance-based offloading and energy-based offloading.

- Incremental Cryptography:** The solution of storage capability of mobile devices is that the energy-efficient protocol that is used for the merging of storage service in MCC. Here, Itani et al. [1] proposed the protocol that uses the concept of incremental cryptography and confidence system which uses data structure that actually secure user data, while the energy consumption of mobile user reduces



**Fig. 1** Categorization of energy efficiency problem in MCC

if it supports the dynamic data operation. System model has three main elements that is mobile client that uses the services of cloud, cloud service provider that manages the cloud services, and trusted third party that is responsible for provide communication between mobile devices and the cloud server.

Crypto coprocessor is allocated to the multiple register client and shares each individual client the unique secrete key  $K_s$ . The system computation is distributed into three operations that are initialization phase, data verification phase, and the data update phase.

**Initialization phase**—In this phase, data is prepared with the incremental integration code before sending to the cloud.

**Verification phase**—In this phase, mobile client can send request for the verification of file, collection of file, or the whole file system stored in the remote cloud server.

**Data Update Phase**—This phase states that how the incremental cryptography concept can be used to provide security and how we can efficiently perform dynamic operation on cloud data.

2. **Offloading Techniques:** Guo et al. [3] introduced that there are various offloading policies that are discussed in this literature. These policies can be classified as **(i) performance-based offloading policies and (ii) energy-based offloading policies.**

**Performance-Based Offloading:** The aim of performance-based offloading is to enhance the performance of mobile devices in terms of maximizing the execution and completion time and enhance the throughput by using cloud resources, and resource-rich computations are offloaded to the cloud. Satyanarayanan et al. [4] gave their model that uses the virtual machine that works on trusted system and computers which are rich of resources and a group of computers which is called cloudlet. Yang et al. [5, 6] studied that multi-user that performs the computation is to be partitioned so that we have to optimize the data stream of application such that application has to give the maximum throughput.

**Energy Based Offloading:** In energy-based offloading, goal is to reduce the energy consumption of mobile devices by offloading the heavy computation task to the cloud. Here, author uses the straightforward offloading decision based on order to use the computation to communication ratio. Huang et al. [7] in his paper set an application completion deadline and gave the dynamic offloading algorithm which is helpful to preserve the energy of mobile devices in such a way so that we have to achieve the joint optimization of offloading decision and clock frequency control.

### **Energy Efficiency of Mobile Devices**

Energy efficiency of mobile device is one of the major issues in MCC; there are many solutions provided by many researchers where we had focused on four major solutions as resource efficiency, dynamic voltage frequency and scaling (DVFS), process level and system level offloading, and branch and bound algorithm.

- 1. Resource Efficiency:** In order to save the energy, we have to perform offloading of data from smart mobile devices to the cloud, but sometimes offloading of data is not always energy preserving as compared to local mobile device computations. Hence, strategy has been designed that decides whether to offload the data or not. This strategy mainly focuses on the energy efficiency (EE) and spectral efficiency (SE) so that we have to utilize the resources that we have. New paradigm that is called as EE-SE tradeoff, namely resource efficiency (RE) which combines EE and SE in orthogonal frequency division multiplexing (OFDM) which is proposed by Niu et al. [2].

The data transmission between the mobile devices and cloud server is assumed to be taken in OFDM system. OFDM uses all the links with different spectral frequencies. Let the total network width is partitioned into  $k$ th subcarrier. User wants to execute the application, and we have to decide where the application is to be executed that is either in the mobile devices or in the cloud server. Once the application is to be executed, then we have to decide that whether to migrate or not; if we offload the data, then our main focus is on energy efficiency of mobile devices.

The tradeoff between EE and SE can be understood in the following manner. In OFDM network, the total bandwidth  $\mathbf{B}$  is divided into  $k$ th subcarrier and each of bandwidth  $\mathbf{BC} = \mathbf{B}/\mathbf{K}$ . The transmitter is sent a block of bit of size  $s$ . The maximum data rate which is to be achieve for  $k$ th subcarrier is represented as

$$Z_k = B_c \log_2 (1 + P_k |H_k|^2 / N_0 B_c) \quad (1)$$

where  $P_k$  represents the transmission power and  $H_k$  represents the channel frequency.  $N_0$  is the single-sided buzz spectral density. Hence, the overall task completed per unit time  $\mathbf{R}$  and total transmission power  $\mathbf{P}_T$  can be described as follows:

$$\begin{aligned} \mathbf{R} &= \sum_{k=1}^Q Z_k \\ \mathbf{P}_T &= \sum_{k=1}^Q P_k \end{aligned} \quad (2)$$

EE and SE are to be optimized in such a way so that we have to use the resources more efficiently. The two objectives are merged into one objective, which are as follows:

$$\max \beta_1 \eta_{\text{EE}} + \beta_2 \eta_{\text{SE}}$$

where  $\beta_1 + \beta_2 = 1$ ,

$$\beta_1 > 0, \quad \beta_2 > 0 \quad (3)$$

where  $\beta_1$  and  $\beta_2$  are constants.

Simplifying the above formula and deriving RE as follows:

$$\begin{aligned}
 \eta_{RE} &= \eta_{EE} + \alpha\eta_{SE} \\
 &= \eta_{EE} + \alpha(\mathbf{B}_t/\mathbf{P}_t)\eta_{SE} \\
 &= \eta_{EE}(1 + \alpha(\mathbf{B}_t/\mathbf{P}_t)\eta_{SE}/\eta_{EE}) \\
 &= \eta_{EE}(1 + \alpha(\mathbf{B}_t/\mathbf{P}_t)\mathbf{P}_T/\mathbf{B}) \\
 &= \eta_{EE}(1 + \alpha\eta_P/\eta_W)
 \end{aligned} \tag{4}$$

where  $\eta_p$  and  $\eta_w$  represent power utilization and bandwidth utilization, respectively, which is given as:

$$\eta_P = \frac{\mathbf{P}_T}{\mathbf{P}_t}, \quad \eta_W = \frac{\mathbf{B}}{\mathbf{B}_t} \tag{5}$$

where  $\mathbf{B}_t$  is the total bandwidth,  $\mathbf{P}_t$  is the overall power budget, and **alpha( $\alpha$ )** is the weighted factor to control the balance of EE and SE.

Alpha is used to identify, if the optimization problem which will focus on EE and SE. If the transmission power  $\mathbf{P}_T$  is numerically smaller than the bandwidth  $\mathbf{B}$ , then energy efficiency is larger than spectral efficiency. As such  $\mathbf{B}$  and  $\mathbf{P}$  which also act as the unit normalizer balance EE and SE in the numerical simulation.

$$\begin{aligned}
 &\max \sum_{K=1}^Q Z_K \left( 1 + \alpha \frac{\mathbf{P}_T/\mathbf{P}_t}{\mathbf{B}/\mathbf{B}_t} / (\mathbf{P}_T + \mathbf{P}_C) \right) \\
 &\sum_{K=1}^Q Z_K \geq \mathbf{R}_{min} \\
 &\sum_{K=1}^Q Z_K \leq \mathbf{P}_{max}
 \end{aligned} \tag{6}$$

where  $\mathbf{R}_{min}$  is the required minimum rate for the network and  $\mathbf{P}_{max}$  is the maximum allowable transmission power.

In Optimal Resource Allocation Schemes the transmission power and occupied bandwidth resource, it is needed to be optimize to get the maximum RE (Resource Efficiency). We have to determine that RE initially strictly decreases and then strictly increases and then strictly decreases with  $Q$  for a fixed transmission power  $\mathbf{P}_T$ . In order to obtain and drive the optimal solution, we drive the following equations by fixing transmission power  $\mathbf{P}_T$  and active subcarriers  $Q$ .

For a fixed transmission power  $\mathbf{P}_T$ , with the uniform power allocation scheme, the maximum achievable RE at a certain bandwidth  $\mathbf{B}$  can be defined as given below

$$\eta_{RE} = \max \sum_K^Q Z_K \left( 1 + \alpha \frac{\mathbf{P}_T/\mathbf{P}_t}{\mathbf{B}/\mathbf{B}_t} / (\mathbf{P}_T + \mathbf{P}_C) \right)$$

$$\mathbf{P}_k = \frac{\mathbf{P}_T}{Q}$$

$$\mathbf{Q} \times \mathbf{B}_C = \mathbf{B} \quad (7)$$

For energy consumption model, if we have to perform the computation that requires  $C$  numeral of instructions. The speed of the cloud and mobile devices are represented by  $S$  and  $M$ , respectively. The same operation takes  $C/S$  seconds to complete the task at server, and it takes  $C/M$  seconds to complete the task at the mobile devices. If  $U$  is the transmission capacity, cloud and mobile devices interchange  $D$  bytes of data and the data which is transmitted and receive it takes  $D/U$  seconds. At the time of computation, mobile devices take  $\mathbf{P}_c$  for computation,  $\mathbf{P}_i$  when it is in ideal state, and  $\mathbf{P}_T$  for the transmitting and receiving of data. When the mobile devices migrate the task to the cloud, then the amount of power that we conserved is as follows:

$$\mathbf{E}_{\text{save}} = \mathbf{P}_c * (C/M) - \mathbf{P}_i * (C/S) - \mathbf{P}_T * (D/U) \quad (8)$$

where  $\mathbf{P}_c$  and  $\mathbf{P}_i$  are to be confirmed by mobile devices and  $\mathbf{P}_T$  can be obtained at the time of resource allocation.

The optimal transmit power  $\mathbf{P}_T$  and occupied bandwidth  $\mathbf{U} = \mathbf{Q} \times \mathbf{W}_C$ . We put this in equation Eq. 8 in order to determine how much energy is conserved by the mobile devices that are calculated as  $\mathbf{E}_{\text{save}}$  when offloading. The energy conserved is actually a positive number. The values of  $M$ ,  $P_i$ , and  $P_c$  are parameters specific to the mobile system. Hence, when  $\mathbf{E}_{\text{save}}$  is positive, we choose the offloading to calculate; otherwise, we choose to calculate on the mobile device. Through this strategy, we could conserve as much resource consumption as possible.

2. **Dynamic Voltage Frequency Scaling (DVFS):** Patricia Arroba et al. [8] introduced how to solve the problem of power management of cloud server. Dynamic voltage and frequency scaling (DVFS) helps to reduce the usage of resources dynamically which is not used in an efficient manner. If this strategy decreases significantly, then in that case we have to use the static consumption by removing the active server to increase their utilization.

Here, basically we have to use energy optimization strategy for cloud data center, which is a combined form of DVFS and consolidation technique. This guideline is not only known the server for the utilization of the incoming workload but also aware for the impact of its allocation in terms of frequency. One of the main threats for designing of the data center optimization is that how we can implement the fastest algorithm that can process the data during run time. So the research is mainly focused on design of an optimization algorithm which is easy to use in terms of computation requirement, in which both its decision making and the execution in real infrastructure are fast. The algorithm which is to be proposed here is based on a bin packing problem where server acts as a bin with certain variable size due

to the frequency scaling. To design optimization techniques, we first describe the performance and the power contribution in terms of architectural parameter which is influenced by DVFS.

3. **Process level and system level offloading:** Boukerche et al. [9] introduced that there are two types of offloading such as process level offloading optimization and system level. MCC architecture is analyzed in such a way so that we have to reduce the overhead of energy cost and to improve the execution efficiency. Due to the allotted nature of MCC, overhead is obvious because cloud can perform computation corresponding to the mobile client request and it processes large amount of request probably at the same time.

In MCC task offloading, we have assumed that the cloud can create duplicate copies of the mobile client task and process it when mobile client needed; they suggest to use the process level virtualization technology that is used to run the application on the server exactly same that it runs on the mobile devices, it is similar to the JVM (Java Virtual Machine) that uses to run the Java programs or application in different operating systems like Windows, MAC OS, Solaris OS, Linux, and Unix. It is used to reduce the energy consumption of mobile devices.

We mainly focus on the communication delay and the mobile device energy efficiency. That is based on the client–server model via TCP (Transmission Control Protocol) the number of hops in between the Mobile Client and Server can increase propagation delay, queuing delay and processing delay that increases the amount of energy consumption which is use during transmission of data. So the result of this paper states that the Wide Area Network (WAN) network with multiple hops cannot adequately provide the proper network performance to fulfill the offloading requirement when we scaling up. So it suggests the solution of Virtual Machine (VM) that the computing instance preparation overhead can be significantly reduced which is also helpful to decrease the task queuing delay.

Zhang et al. [10] analyzed that the offloading of application to the cloud is energy saving or not, so the analysis was made by the tradeoff between the computation energy for mobile execution and communication energy under fix network for cloud execution.

### **Limited Battery Lifetime of Mobile Devices**

Lifetime of battery is one of the major issues and challenges for mobile devices. There are many solutions provided where two we are discussing in our paper as static probability and energy-efficient computational offloading framework (EECOF).

1. **Static Partitioning:** In MCC, mobile devices are not as much as powerful as desktop because they process or run higher configurable application which is not possible in mobile devices due to its limited battery life, so to enhance the battery lifetime of mobile devices, we have to offload the data to the cloud, but some time it is not convenient because of communication overhead and they consume large amount of energy, so we favor local computation.

Saab et al. [11] introduced the concept which is called as static partitioning, in which we divided the mobile data into small chunks of packets and then

dynamically offloaded it to the cloud, and some time it is not energy-efficient, but it reduces the computation time of data.

Here, it uses the clone cloud architecture in which we have to clone the smartphone in cloud so that we partially offload the data to the clone cloud and then offload it to the main cloud in such a way that we reduce the energy consumption of mobile devices and enhance battery lifetime.

MCC uses to save the energy of mobile users by offloading the application to the cloud, but we are not able to offload the entire data to the cloud at a time because of the power limitation of cloud. Cloudlet concept is used to improve the performance of MCC. It considers the type of network and the distance from mobile devices to the cloud server that put impact on the power consumption of offloading of application.

The Cloudlet-based rich resources which use virtual machine which is near to the Cloudlets to improve the application management, and we are able to process large amount of application compared to the application that we process before. Factors that influence the energy consumption of mobile devices are categorized into mobile handset, workload, and the communication network:

- **Mobile Handset**—Each and every mobile handset has their own power let say ( $P_n$ ), and it performs certain number of instruction which we call computations ( $C_n$ ) and the power use the mobile handset when it is idle ( $P_{idle}$ ). The power used by the mobile handset for transmitting and receiving particular amount of data over network is represented by ( $P_{se}$ ) and ( $P_{re}$ ), respectively. So we are known these values by the mobile handset and also about the communication network.
- **Workload**—Decisiveness of division totally depends upon the amount and types of workload that we have. The same application with different amount of data leads to two different decisions. So the amount of processing required to every module ( $C_n$ ) of data that send to the cloud ( $D_{se}$ ) and received ( $D_{re}$ ).
- **Communications Network**—In communication network, the rate of bits for sending ( $B_{se}$ ) and receiving ( $B_{re}$ ) are the factors that influence the speed of transmission between mobile devices and the cloud server. If there is delay in the communication, it results in the enhancement of power consumption. So to reduce the power consumption, we use the different types of network like 3G, 4G LTE.

So in this paper, author introduces the **static partitioning** for mobile application that if we offload the application to the cloud after partitioning, then it reduces the power consumption of mobile devices and enhances their battery lifetime and if there is light-weighted application so no need to offload it to the cloud.

2. **Energy-Efficient Computational Offloading Framework (EECOF)**: Vinh et al. [12] presented the concept on the energy efficiency of MCC that leads to reduce the battery capacity of mobile devices, where the computation resources in mobile devices are less, so there is need of offloading the application to the environment where resources are more, like cloud and after process the application in the cloud results send to the mobile devices again to reduce energy consumption

and enhance the battery lifetime of mobile devices. They proposed a distributed **Energy-Efficient Computational Offloading Framework** (EECOF) for MCC. This framework offloads the computation deep application to the cloud at the run time. The resultant is to reduce the data transmission rate and power efficiency of mobile devices while offloading the application to the cloud. However, it seems to be bandwidth consuming when we require to transfer the data from mobile devices to the cloud.

Dynamic **resource provisioning scheduler** that is used to offload the data to the cloud that minimizes the energy consumption over computation and communication. The scheduling can perform by the internet based virtual data for adaptive resource management. In smart mobile devices energy consumption increases when it try for face finder and in order to consider that the delay of data transmission is always constant is not possible in real world setting. They proposed free sequence protocol permits dynamic execution of application on the cloud server. It actually uses the compression technique to offload the data to the cloud. They found that Wi-fi is better than 3G because 3G consume more battery and network bandwidth than Wi-fi.

3. **Application Layer Adaptive Transmission Protocol:** Liu et al. [13] proposed AppATP (Application Layer Adaptive Transmission Protocol) whose aim is to provide energy-efficient transmission between mobile devices and the cloud. The need of AppATP can be observed by the following two observations that is—first, the energy consumption in transmission is affected due to the unknown nature of the wireless network; it is because of the fluctuation in wireless network. Scaling shows that large amount of energy is consumed during poor connectivity and less amount of energy consumed during better connectivity in mobile devices. Second, many mobile applications are delay tolerant like YouTube, Netflix that is why we require AppATP to provide energy-efficient transmission between mobile devices and the cloud. AppATP consumes the cloud to provide cloud resources in order to give the transmission management. It buffers the data of application in cloud before transmitting them to the mobile devices in order to schedule the transmission data of mobile app. This is best suitable for large range of delay-tolerant. In such a way, we can resolve the problem of limited battery lifetime of mobile devices.

### 3 Conclusion

This paper focuses on various energy efficiency issues in MCC and solutions for the same. The issue for efficient efficiency in MCC is divided into three major issues as storage capacity, energy efficiency, and limited battery lifetime. The storage capacity is issue related to limited storage capacity of small mobile devices. There are two solution approaches discussed in this paper as incremental cryptography and offloading techniques to overcome with this problem. Energy efficiency is another challenge in MCC which requires optimized utilization of energy. Solution approaches discussed

in this paper for energy efficiency are resource efficiency, DVFS approach, process level and system level offloading, and branch and bound algorithm. Next major issue of MCC discussed is limited lifetime of battery which is due to small size of mobile phones. Various solutions discussed for these issues are static partitioning and EECOF framework which maximizes the lifetime of battery for MCC. Thus, this paper proposes a survey for different energy efficiency problems and their solutions for MCC.

## References

1. Itani W, Kayssi A, Chehab A (2010) Energy-efficient incremental integrity for securing storage in mobile cloud computing. IEEE
2. Niu C, Yang S, Wang F (2015) A unified energy efficiency and spectral efficiency tradeoff for mobile cloud computing in OFDM-based networks. IEEE, pp 306–311
3. Guo S, Liu J, Yang Y, Xiao B, Li Z (2018) Energy-efficient dynamic computation offloading and cooperative task scheduling in mobile cloud computing. IEEE
4. Satyanarayanan M, Bahl P, Caceres R, Davies N (2009) The case for VM based cloudlets in mobile computing. IEEE Pervasive Comput 8(4):14–23
5. Yang L, Cao J, Cheng H, Ji Y (2015) Multi-user computation partitioning for latency sensitive mobile cloud applications. IEEE Trans Comput 64(8):2253–2266
6. Yang L, Cao J, Tang S, Li T, Chan ATS (2013) A framework for partitioning and execution of data stream applications in mobile cloud computing. ACM SIGMETRICS Perform Eval Rev 40(4):23–32
7. Huang D, Wang P, Niyato D (2012) A dynamic offloading algorithm for mobile computing. IEEE Trans Wireless Commun 11(6):1991–1995
8. Arroba P, Moya JM, Ayala JL, Buyya R (2015) DVFS-aware consolidation for energy-efficient clouds. IEEE, pp 494–495
9. Boukerche A, Guan S, De Grande RE (2018) A task-centric mobile cloud-based system to enable energy-aware efficient offloading. IEEE
10. Zhang W, Wen Y (2015) Energy-efficient task execution for application as a general topology in mobile cloud computing. IEEE
11. Saab SA, Chehab A, Kayssi A (2013) Energy efficiency in mobile cloud computing total offloading selectively works. Does selective offloading totally work? IEEE, pp 164–168
12. Vinh TL, Pallavali R, Houacine F, Bouzefrane S (2016) Energy efficiency in mobile cloud computing architectures. IEEE, pp 327–331
13. Liu F, Shu P, Lui JCS (2015) “AppATP: an energy conserving adaptive mobile-cloud transmission protocol. IEEE

# Perspectives of Blockchain in the Education Sector Pertaining to the Student's Records



Poonam Verma and Ankur Dumka

**Abstract** Blockchain has been enforced in many sectors, and its implementation has drastically improved those sectors. Diamond trade has been greatly helped by the employment of blockchain to digitally track the diamonds being well-mined and sold-out within the market. The potential of the blockchain is being tested by varied government agencies for distribution of the various services. Through this paper, we've explored the entities concerned within the blockchain network enforced in an academic institute. This paper focuses the employment of the blockchain account for each student maintaining the records of the scholars that helps to spot the talents of the scholars by their prospective employers.

**Keywords** Blockchain · Security · Education · Students · E-certificates · Degrees

## 1 Introduction

The booming era of BitCoin has bought blockchain to one of the topmost trends even beyond the trend of the Deep Learning. Blockchain is based on the distributed decentralized platform that may be used for computation and knowledge sharing. Blockchain will work, coordinate between the multiple authoritative domains that don't trust one another. The distributed decentralized platform helps to rearrange for multiple points of coordination. Blockchain can provide a sturdy consistency support to the members connected to the blockchain. Blockchain helps to keep the transactions secure by employing a cryptographically secure functions. However, blockchain also offers the facility to record the transactions between the permissioned and unpermissioned 2 entities not capable to trust one another in an economical, verifiable, and permanent manner [1]. Blockchain can help to trace the origin and

---

P. Verma (✉) · A. Dumka  
Graphic Era Deemed to be University, Dehradun, India  
e-mail: [poonamddn18@gmail.com](mailto:poonamddn18@gmail.com)

A. Dumka  
e-mail: [ankurdumka2@gmail.com](mailto:ankurdumka2@gmail.com)

ensuant movement of the luxurious merchandise across the provision chain [2, 3]. Once the item is formed at the origin, a corresponding digital token is issued by the trusty entity, which can be the purpose of validation and authentication at the purpose of origin [4]. Every time, the merchandise moves within the supply chain step by step, the digital token can mirror its chains of transactions on the blockchain [5]. The digital token would act as a certificate of credibility, which might be more durable to forge than different entities.

## 2 Design of the Blockchain

Blockchain is created from blocks that are immutable. These blocks consist of the data which is available to the public openly and everybody possibly can validate these transactions within the block [6]. The data within the blocks are going to be digitally signed, and therefore, the transactions in these blocks are to be verified by the peers. The participants will read the data on the ledger that are approved to check. Transactions are organized in an exceedingly block known as a Merkel Tree. Merkel Tree is employed to construct the Block Hash.

Blockchain network has primarily 2 classes of models to cater to and that they are permissioned model and permissionless model [7]. The permissioned model is appropriate for business applications wherever strict protocols are to be enforced like good contracts whereas the permissionless models are additional applicable for open management applications. In this paper, for the education sector, we have considered only the permissioned model.

## 3 Distributed Classic Consensus

Partial synchrony in a very distributed system lies between the cases of a synchronous system and an asynchronous system. In an exceedingly synchronous system, there's a upper limit that is needed for the message to be sent from one processor to a different and a upper limit is additionally obtainable on the relative speeds of various processors [8]. In asynchronous system, no higher bounds exist for the messages to be processed and therefore the speeds of the processors. Fault-tolerant agreement protocols are given for varied cases of partial synchrony and varied fault models. Lower bounds that show in most cases that our protocols are optimum with relevancy the amount of faults tolerated are given. Our agreement protocols for partly synchronous processors use new protocols for fault-tolerant “distributed clocks” that permit partly synchronous processors to achieve some or so common notion of your time [9, 10].

Malicious attacks and package errors are progressively common. The growing reliance of trade and government on on-line info services makes malicious attacks additional engaging and makes the implications of thriving attacks additionally

serious [11]. The amount of package errors is increasing which is due to the expansion in size and complexity of package. Since malicious attacks and package errors will cause faulty nodes to exhibit Byzantine (i.e., arbitrary) behavior, Byzantine-fault-tolerant algorithms are progressively vital. The TolerateByzantine rule works in asynchronous systems, and it incorporates vital optimizations that change it to perform with efficiency [12].

Some of the agreement algorithms by that the blockchain network aims to realize the distributed agreement are listed as below:

- PoW: Proof\_of\_Work is a consensus mechanism that prevents the service attacks and other services abuses by requiring a complicated computational problems to be solved by the service requester [13, 14].
- PoS: Proof\_of\_Stake is a popular consensus algorithm that aims to achieve the distributed level by permitting the miner with the most number of resources to solve the mining problem.
- PoB: Proof\_of\_Burn is a consensus algorithm that permits those miners to participate in the mining only if the number of resources utilized by them is more than the others.
- PoET: Proof\_of\_period\_of\_time is a mechanism that prevents high utilization of the resources and high consumption of energy.

## 4 Cryptography in Blockchain

Prime numbers in randomness are popularly utilized in the keys for several security protocols. Public randomness has been researched to develop a frenzied methodology that may be additional with efficiency be used for higher security against manipulation. Rabin had way back projected a trusty service named as a beacon that may facilitate to broadcast random numbers at regular intervals [12, 15].

Bitcoin miners maintain the block chain, a public organization serving as a world ledger of all transactions within the history of the system [11]. New batches of transactions are revealed in an exceedingly block or so each ten minutes. Any party will make an attempt to publish future valid block of transactions; however, doing therefore is computationally troublesome owning to the proof-of-work system in Bitcoin. An Appropriate block should have a hash worth beginning with n no. of consecutive zeroes. The problem parameter n is regularly adjusted to keep up the typical mining rate of one block per ten minutes. At the time of this writing  $n \approx 66.4$ . Every block contains a hash of the previous valid block. The set of all reputed blocks so forms a tree, with the longest chain because the one valid block chain process current possession of all bitcoins [16, 17].

Some of the distributed cryptanalytic protocols in an exceedingly absolutely peer-to-peer state of affairs projected earlier underneath the idea that the computers have restricted computing power.

1. A broadcast protocol secure underneath the idea that the honest parties have computing power.
2. A protocol for characteristic a collection of parties such the bulk of them is honest, and each honest party belongs to the current set [11].

Human dignity demands that private info, like medical and knowledge remains secret from the overall public, but the new procedure tools are quite useful for the preservation of the privacy, thereby discarding away the trust problems with the organizations. The thought of the zero information proof systems will not want to preserve the non-public privacy and institutional integrity with the assistance of the cryptanalytic algorithms. In most of the social media, the general public demands the transparency from the zero information systems forward that none of their knowledge is unbroken by the organizations for his or her industrial uses. So such systems are being wide used for the crypto knowledge and exponential verification of the information [18].

## 5 Cyber Attacks

The Bitcoin crypto currency records its transactions in an exceedingly public log referred to as the blockchain. Its security rests critically on the distributed protocol that maintains the blockchain, pass by participants referred to as miners. Typical knowledge asserts that the protocol is incentive-compatible and secure against colluding minority teams, i.e., it incentivizes miners to follow the protocol as prescribed. We have a tendency to show that the Bitcoin protocol isn't incentive-compatible [19]. We have a tendency to gift associate degree attack with that colluding miners get a revenue larger than their justifiable share. This attack will have vital consequences for Bitcoin: Rational miners can choose to be a part of the group of miners, and therefore, the colluding cluster can increase in size till it becomes a majority. At this time, the Bitcoin system ceases to be a redistributed currency.

## 6 Use of Blockchain in Education Sector

The use of the blockchain in education is in an infancy phase while giving a series of recommendations to encourage the event of the blockchain within the education sector. Blockchain will act because the hub for confirmative associate degree verificatory the transactions of an institute wherever the users are workers, students, teachers, management employees except for the accreditors, validators, and testers.

Blockchain can help in various ways. In India, the student population is quite huge and to maintain the records of such huge population is a cumbersome process apart from the challenges of the fraud in the distribution of the fake degrees or certificates to the students, thereby wasting the potential of the students. Similarly, the employees

and teachers of the institute are also quite exponential in numbers and to keep up the procedures from their accomplishment to their teaching methodology and to their performance metrics ensuing into their appraisals is additionally a troublesome task.

Another issue to handle the records of the nationwide institutes by the accreditors and auditors with their suitable remarks are also hard to address and store with security. Thus, the above challenges indicate the available scope for the blockchain in the education sector.

## 7 Blockchain Being Used in the Student Challenges

1. Degrees and e-certificates
2. Secure account for the skill certificates
3. Accreditation credibility by the reviewers based on the regular assignments
4. Reliable online transactions of the fees, scholarships, and rewards
5. Fraud detection in distribution of the fake degrees
6. Secure student records.

Blockchain is considered to be an immutable distributed ledger that can maintain the transactions securely. An educational institute with the admission of the students in their institute can set up a record that can store the transactions verified by the public and accreditors. Each individual account is recognized as the unique block with an encrypted key that is issued to the student by the certifying agency. In this paper, we will consider the permissioned scheme of the blockchain where the entities are given permission rights to verify and observe the community transactions.

Once the student starts their studies, they are given regular assignments and projects to assess their skills learned by them during the course. These marks are the transactions that are listed in the student's account. These transactions or records can be seen by the permissioned entities of the minimal network consisting of the nodes of the students, admins, and the certifying agency.

These accounts with the records of each student will also help to provide the insight to the psychological aspects of the students throughout their career path. Each record is verified by the permissioned entities of the network. Block chain can also help in sharing the records of the students more aptly with an appropriate edge of the computational security and privacy. Specific skills of the students can be verified and communicated with the use of the badges or tokens that are again verified by the community and can be shared with the prospective employers of the students.

## 8 Conclusion

In this paper, we have described an overview of the various aspects of the blockchain with its features. We have also specifically chosen the use of the blockchain in the

education sector which benefits the student and teacher community. Blockchain is one of the upcoming trends that can be adopted to provide a reliable secure transactions for the student records. This can be further extended for the teachers and employees as well as the affiliations from various agencies can help to curb the corruption from the nation resulting into a shining Bharat.

## References

1. Westerkamp M, Victor F, Küpper A (2018) Blockchain-based supply chain traceability: token recipes model manufacturing processes. In: 2018 IEEE international conference on internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData), Halifax, NS, Canada, 2018, pp 1595–1602
2. Notheisen B, Hawlitschek F, Weinhardt C (2017) Breaking down the blockchain hype—towards a blockchain market engineering approach
3. Xu X, Pautasso C, Zhu L, Gramoli V, Ponomare VA, Tran AB, Chen S (2016) The blockchain as a software connector. In: 2016 13th working IEEE/IFIP conference on software architecture (WICSA), Apr 2016. IEEE, pp 182–191
4. Sutton A, Samavi R (2017) Blockchain enabled privacy audit logs. In: International semantic web conference, Oct 2017. Springer, Cham, pp 645–660
5. Notheisen B, Cholewa JB, Shanmugam AP (2017) Trading real-world assets on blockchain. Bus Inf Syst Eng 59(6):425–440
6. Ølnes Maupin J (2017) The G20 countries should engage with blockchain technologies to build an inclusive, transparent, and accountable digital economy for all (No. 2017-48). Economics Discussion Papers
7. Shafagh H, Burkhalter L, Hithnawi A, Duquennoy S (2017) Towards blockchain-based auditable storage and sharing of IoT data. In: Proceedings of the 2017 on cloud computing security workshop, Nov 2017. ACM, pp 45–50
8. Castro M, Liskov B (1999) Practical Byzantine fault tolerance. In OSDI, vol 99, no 1999, pp 173–186
9. Anjum A, Sporny M, Sill A (2017) Blockchain standards for compliance and trust. IEEE Cloud Comput 4(4):84–90
10. Badertscher C, Maurer U, Tschudi D, Zikas V (2017) Bitcoin as a transaction ledger: a composable treatment. In: Annual international cryptology conference, Aug 2017. Springer, Cham, pp 324–356
11. Kumar S, Singhal A, Dumka A (2019) Analysis of cloud security using blockchain technology. Selected in ICAESMT19 (International conference on advances in engineering science management & technology 2019) will be held at Uttarakhand University, Dehradun, Uttarakhand, India on 14–15 Mar 2019
12. Beck R, Stenum Czepluch J, Lollike N, Malone S (2016) Blockchain—the gateway to trust-free cryptographic transactions
13. Lamport L, Shostak R, Pease M (1982) The Byzantine generals problem. ACM Trans Prog Lang Syst (TOPLAS) 4(3):382–401
14. Lamport LB (2010) U.S. Patent No. 7,711,825. Washington, DC: U.S. Patent and Trademark Office
15. Andrychowicz M, Dziembowski S (2014) Distributed cryptography based on the proofs of work. IACR Cryptology ePrint Archive, 2014, p 796
16. Alharby M, van Moorsel A (2017) Blockchain-based smart contracts: a systematic mapping study. arXiv preprint [arXiv:1710.06372](https://arxiv.org/abs/1710.06372)

17. Dolev D, Lamport L, Pease M, Shostak R (1987) The Byzantine generals. In: Concurrency control and reliability in distributed systems. Van Nostrand Reinhold Co., pp 348–369
18. Maupin J (2017) The G20 countries should engage with blockchain technologies to build an inclusive, transparent, and accountable digital economy for all (No. 2017-48). Economics Discussion Papers
19. Bonneau J, Clark J, Goldfeder S (2015) On Bitcoin as a public randomness source. IACR Cryptology ePrint Archive, 2015, p 1015

# Ground-Level Water Predication Using Time Series Statistical Model



Sandeep Kumar Mittal, Mamta Mittal, and Muhammad Sajjad Ali Khan

**Abstract** With exponential increase in population, scarcity of water in Nation Capital of India, Delhi, has become the most critical issue over last few years. The change in climatic conditions, usage of lands and immense abstraction of water are plausible reasons for depletion of groundwater at a rapid rate. To deal with this issue, authors predict the groundwater level using time series model in various regions of Delhi. To understand the reasons of decline in groundwater level and its quality is necessary for the development and livelihood in all regions of Delhi. The results depict that decline in number of wells from 125 in year 2012 to 82 in the current year. Over the period of six years, lower rain falls and high population growth is the major reasons for depletion of groundwater level in Delhi. Along with this, the quality of ground water has been deteriorated which has also become a prime issue of concern.

**Keywords** Groundwater level · Time series analysis · Predication · NCR Delhi

## 1 Introduction

Ground water is the largest critical source of fresh water that is world-wide available. This essential water resource is susceptible to change in climatic conditions and extraction at untenable rates [1, 2]. This crucial resource is replenishable and dynamic in nature. Globally, 38% of the groundwater is used for irrigation of agricultural lands

---

S. K. Mittal

Department of Mathematics, G. B. Pant Government Engineering College, New Delhi, India  
e-mail: [mittalsandeep1983@gbpec.edu.in](mailto:mittalsandeep1983@gbpec.edu.in)

M. Mittal (✉)

Department of CSE, G. B. Pant Government Engineering College, New Delhi, India  
e-mail: [mittalmamta79@gmail.com](mailto:mittalmamta79@gmail.com)

M. S. A. Khan

Department of Mathematics, Institute of Numerical Sciences, Kohat University of Science and Technology, Kohat, Khyber Pakhtunkhwa, Pakistan  
e-mail: [sajjadali.math@yahoo.com](mailto:sajjadali.math@yahoo.com)

[3]. It is also one of the major sources of drinking water in urban and rural parts of India, and about 33% of groundwater is supplied for household purposes. Also, it is of utmost importance in the field of agriculture and industrial sectors. Being a critical natural resource of water, its availability depends on the rainfall and restoration conditions and it is considered as most dependable source of uncontaminated water ([edugreen.teri.res.in](http://edugreen.teri.res.in)).

To fulfil the requirements of livelihood, large population purely relies on the groundwater. With increasing rate of population, recent studies have detected depletion of groundwater resources at a swift rate worldwide through satellite observations [4–6]. For assessing its availability and to preserve a vital natural resource, it is essential to continuously monitor groundwater level. To predict and monitor the groundwater level, several techniques and methods have been used in previous studies. In literature, S. N. Bhanja et al. measured groundwater level through GRACE satellites [7]. In their consecutive study [8], ground water is predicted through spatial and temporal variability. Apart from these, several machine learning and artificial intelligence techniques such as artificial neural network [9–11], random forest regression [12] and single well and polynomial models [13] are used in literature. Apart from these techniques, some statistical techniques such as time series is used to monitor the natural resources of water. Time series analysis, also referred as trend analysis, deals with the time series data. It is described in terms of two components: trend and seasonality. Trend represents linear and nonlinear component that changes with time but does not repeat itself whereas seasonality represents changes over time and repeating itself over systematic time intervals.

In this paper, authors have used a statistical technique, time series analysis to predict and monitor the depletion of ground water in the region of national capital of India, Delhi. The prediction is done as per various regions in Delhi, i.e., central, east, New Delhi, north, north-east, north-west, south, south-west and west. The data are collected over the period of 2012–2018 [14]. As per records, no publicly data are available for the year 2019. Groundwater level is monitored in terms of number of wells, water below the ground and others. From the results, it was found that water quantity in various wells is decreasing day by day in overall regions of Delhi. It is predicted that water level is very low and it cannot satisfy and fulfil the needs of people for day to day activities. The levels of water in wells are also decreasing, and rainfall alone cannot fulfil the need of water in Delhi region. This is a major issue of concern, and it is desirable to take appropriate measures to prevent water level from depletion.

## 2 Related Work

With the increasing rate of population worldwide, groundwater level is depleting at a rapid rate. Several researchers are predicting and monitoring groundwater level across various parts of the world. Bhanja et al. [7] have compared GRACE-based

GWS (groundwater storage) anomaly on two recent GRACE datasets, RL 05 spherical harmonies (SH) solutions and RL 05 meson solutions (MS) over the period of 2005–2013. It was found that GRACE-based GWS anomalies reveal firm seasonality as compared to GWS satellites. This has been estimated using several statistical techniques such as RMSE, Correlation, skewness, kurtosis and others. Based on results, it was recommended to use GRACE-MS estimates for study of groundwater level. In their consecutive study [8], the authors have studied groundwater level measurements of 3907 wells in 22 rivers basis of India. In this, they have studied the spatiotemporal variability of GWS anomalies. It was concluded that GWS anomalies are influenced by spatial variability, i.e., well spacing. It increases with increase of spatial extent i.e.  $0.25^\circ$ ,  $0.5^\circ$  and  $1^\circ$ . The main advantage of this work is cost-effective design for monitoring ground water. Apart from spatial and satellite prediction, Wang et al. [10] proposed an algorithm named canonical correlation forest algorithm with random features (CCF-CRF). The algorithm had been evaluated on the data of Daguhe River in Oingdao, China. The results depict that CCF-CRF achieves improved performance and computation time as compared to other algorithms. Thus, CCF-CRF is a remarkable prediction tool used for groundwater hydrology. Artificial neural networks are also used to predict and monitor the water level of 15 sites. The results concludes that ANN achieves better values for correlation and thus can be considered as a robust tool for prediction of dynamic groundwater level [9]. Groundwater level fluctuations have been forecasted in an unconfined coastal aquifer in India using artificial neural networks [15]. M. Mirzavand et al. predict the groundwater level in Kashan Plains, Iran, using time series analysis [16]. R. Reghunath et al. estimated long-term fluctuations in water table in river basis using time series [17]. D. C. Neto et al. used several methods of time-series analysis such as Fourier analysis, cross-correlation and R/S analysis to monitor the groundwater table in wells in the duration of 2002–2005 [18].

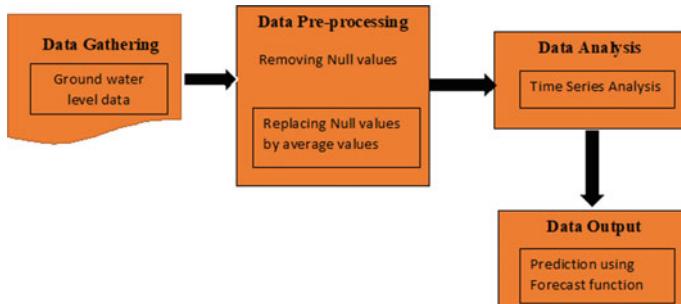
In this research paper, authors used a time series model to predict the ground level water. Time series models predict the future water fall in Delhi region. Main motivation to used time series model is that it works on the small size of data, whereas CNN or deep learning models requires huge data and computation cost is low.

### 3 Methodology

In this section, authors have explained the complete methodology for predication of water level in Delhi region. In the first section, authors have explained the data collection and process flow diagram and next section has a brief about the time series model.

#### 3.1 Process Flow Diagram

Figure 1 depicts the process flow diagram of the proposed approach. The groundwater



**Fig. 1** Process flow diagram

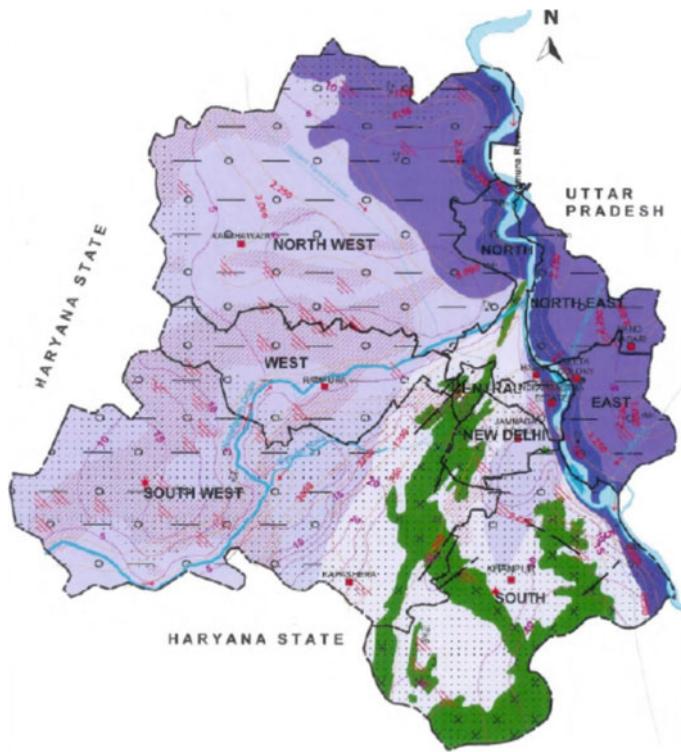
level data of National Capital territory of Delhi is gathered. After data collection, data are preprocessed by removing all the null values and null values are replaced by average values. To predict the groundwater level in various regions of Delhi in future, pre-processed data over the year 2012–2018 is analysed using time series analysis. After data analysis, groundwater level is predicted using Forecast function.

### 3.2 *Data Preparation and Studied Area*

To predict the groundwater level, data of National Capital Territory of India, Delhi, are gathered. Data are collected from the hydrological department of Delhi [19] over the period of 2012–2018. To predict the groundwater level, number of wells and water level in wells are considered as a main parameter of this work. Delhi is further divided into several regions as: East, New Delhi, north, north-east, north-west, south, south-west and west regions as depicted in Fig. 2. The groundwater level varies from season to season, and it increases at the time of rainfall season and decreases after that. The fluctuations in water level of various regions of Delhi are depicted in Fig. 3.

Number of wells in different regions of Delhi was counted from the data gathered by hydrological department. It was found that number of wells was 125 in the year of 2012 which had decreased to 82 in the year of 2018. The number of wells and fluctuations in water level in various regions over the number years is depicted in Fig. 4 and Table 1. To predict these falls in number over the years, time series analysis is used as explained in Sect. 3.3.

Table 1 shows the data of wells present in the year 2015 in various region of Delhi. Similarly, Fig. 4 shows the data of wells in different regions from the year 2012 to 2018.



**Fig. 2** Various regions in Delhi [20]

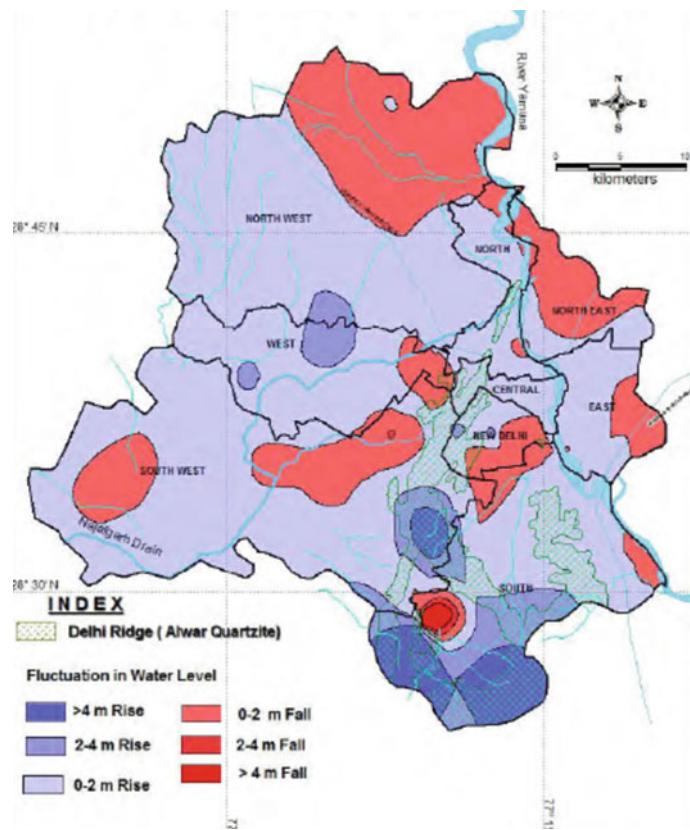
### 3.3 Data Time Series Analysis

Time series analysis is also referred as trend analysis which develops mathematical models to capture an observed time series and analysis of all possible descriptions from the time series data. Time series model can be applied over daily, monthly, quarterly and yearly basis. Four different components of time series which are analysed are as follows: secular trend which observes whether data are linear or non-linear; cyclical variation determines rise or fall over periods; seasonal variation to determine change of patterns; and residual variation to determine noise [14, 19]. It is represented by a mathematical equation in Eq. 1.

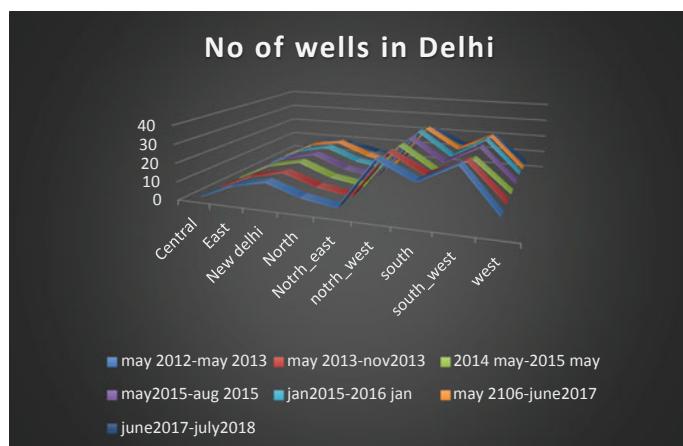
$$Y_t = T_t + C_t + S_t + R_t \quad (1)$$

Time series with the linear trend can be represented in Eq. 2, and this is similar to linear regression model.

$$Y_t = a + bt + e_t \quad (2)$$



**Fig. 3** Fluctuations of water level over various regions of Delhi [21, 22]



**Fig. 4** Number of wells in different regions of Delhi from year 2012 to 2018

**Table 1** Number of wells present in different regions of Delhi

District Name	No of wells	Range of fluctuation (m)				No of well/percent showing fluctuation						Total no of wells	
		Rise		Fall		Rise		Fall		Rise			
		Min	Max	Min	Max	0–2	2–4	>4	0–2	2–4	>4		
Central	1	0.98	0.98	—	—	1 100.00%	0	0	0	0	0	1 0	
East	10	0.03	1.07	0.26	0.27	8 80.00%	0	0	2 20.00%	0	0	8 2	
New Delhi	14	0.23	5.18	—	—	11 78.57%	2 14.29%	1 7.14%	0	0	0	14 0	
North	7	0.56	1.45	—	—	7 100.00%	0	0	0	0	0	7 0	
North-east	4	0.55	0.55	0.14	0.49	1 25.00%	0	0	3 75.00%	0	0	1 3	
North-west	28	0.01	3.12	0.08	2.81	15 53.57%	5 17.86%	0	6 21.43%	2 7.14%	0	20 8	
South	16	0.29	13.42	0.24	9.92	7 43.79%	1 6.25%	4 25.00%	3 18.75%	0	1 6.25%	12 4	
South-west	26	0.14	4.42	0.06	2.10	14 53.89%	2 7.69%	1 3.85%	8 30.77%	1 3.85%	0	17 9	
West	8	0.22	3.68	—	—	4 50.00%	4 50.00%	0	0 0	0	8 0		

In case of non-linear trends, time series models are transformed the linear to logarithmic models (represented in Eq. 3). In the case of non-linear models, if time series is decreased over the time, Eq. 3 can be transformed, represented in Eq. 4.

$$\log(Y_t) = a + bt + e_t \quad (3)$$

$$Y_t = a + b \ln(t) + e_t \quad (4)$$

In non-linear trends, curvilinear trends can also be represented by the second order of polynomial in Eq. 5.

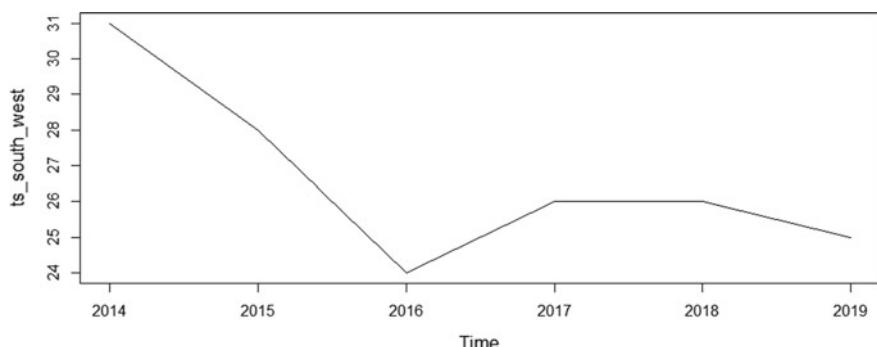
$$Y_t = a + b_1 t + b_2 t^2 + e_t \quad (5)$$

In the time series model, accuracy for forecasting is computed by mean squared error or mean absolute derivation.

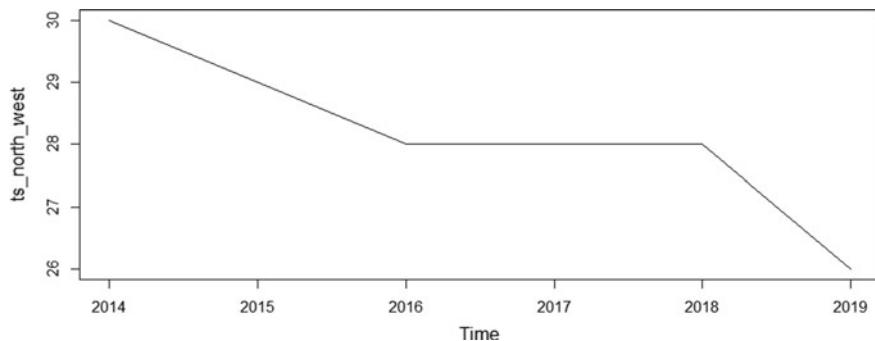
## 4 Results and Discussions

Time series analysis is used to estimate the groundwater level in various regions of Delhi by predicting number of wells over the time period of 2012–2019. Authors have calculated the total numbers of wells in year 2012–2018 in the various regions of Delhi. Using R language, the function `ts()` is used for time series analysis. For illustration, the results are depicted in Figs. 5 and 6 for south-west and north-west regions.

The results depicted that in south-west region, number of wells is decreasing slowly, but in case of north-west region, there is a steep decrease in the number of wells in year 2019. Using this `ts()` function, authors have computed the value of



**Fig. 5** Time series prediction of south-west region



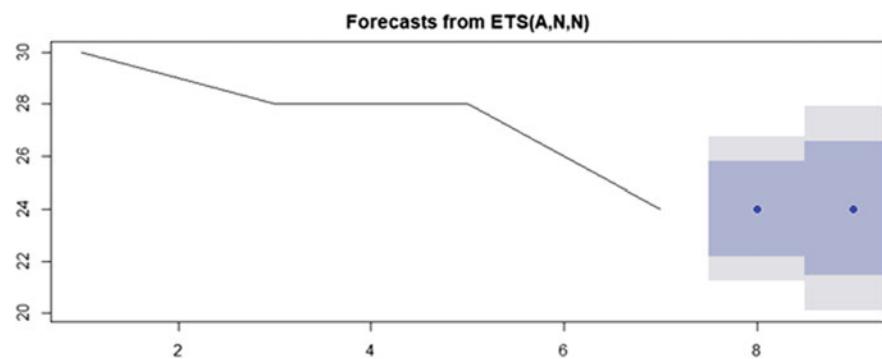
**Fig. 6** Time series prediction of north-west region

wells in different regions of Delhi. Table 2 and Fig. 7 show the forecasted value of south-west and north-west region for next two years.

The total results depict that total number of wells had decreased from 125 to 82 since 2012 to 2018, and in year 2019, it will be 80 due to heavy rain fall in the entire region. But 14% more fall in the quantity of waters in wells as compared to previous year. In case of fluctuations of water level, in 2012, from 125 wells level of water risen in 118 wells and fall in water level is observed in only 11 wells. As compared to 2018, among 82, water level of only 35 wells had risen and rest experience fall in

**Table 2** Forecasted value of south-west and north-west regions

North-west	Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
	8	24.0002	22.18772	25.8126	21.22825	26.77215
	9	24.0002	21.43710	26.56330	20.08027	27.92013
South-west	9	26.37468	22.83087	29.91849	20.95489	31.79447
	10	26.37468	22.83087	29.91849	20.95489	31.79447



**Fig. 7** Forecasted values of north-west regions

water level. Therefore, it can be predicted that water level in Delhi is very low and is not capable of fulfilling the needs of people as it is decreasing day by day. People in Delhi are now only dependent on rain water. This is the serious issue which needs a consideration for future life.

## 5 Conclusion

To improve upon the quality and level of groundwater in Delhi, high yielding aquifers should be used for effective mitigation of drinking water. With tremendous increase of population, one of the major reasons for depletion of ground water, harvesting structures should be improved along with preservation and management of floodplains and heavy flow of water in rivers. Number of wells should also be monitored with time and effective measures should be taken to preserve ground water. To monitor and predict the level of ground water over the period of 2012–2018, time series analysis is used. The results depict that decrease in number of wells from 125 to 80 over the period of six years in various regions of Delhi is one of the major causes for depletion of groundwater level.

## References

1. Wada Y, Van Beek LPH, Van Kempen CM, Reckman JWM, Vasak S, Bierkens MFP (2010) Global depletion of groundwater resources, vol 37, Sept, pp 1–5
2. Taylor RG (2019) Ground water and climate change, Nov 2012
3. Siebert S, Burke J, Faures JM, Frenken K, Hoogeveen J (2010) Groundwater use for irrigation—a global inventory, pp 1863–1880
4. Rodell M, Velicogna I, Famiglietti JS (2009) Satellite-based estimates of groundwater depletion in India. *Nature* 460(7258):999–1002
5. Voss K, Swenson S, Rodell M (2015) Quantifying renewable groundwater stress with GRACE, pp 5217–5238
6. Voss KA, Famiglietti JS, Lo M, De Linage C, Rodell M, Swenson SC (2013) Groundwater depletion in the Middle East from GRACE with implications for transboundary water management in the Tigris-Euphrates-Western Iran region, vol 49, pp 904–914
7. Bhanja SN, Mukherjee A, Saha D, Velicogna I, Famiglietti JS (2016) Validation of GRACE based groundwater storage anomaly using in-situ groundwater level measurements in India. *J Hydrol* 543:729–738
8. Bhanja SN, Rodell M, Li B, Saha D, Mukherjee A (2017) Spatio-temporal variability of groundwater storage in India. *J Hydrol* 544:428–437
9. Johnny C, Sashikumar MC (2015) Prediction of groundwater level dynamics using Artificial Neural Network Prediction of groundwater level dynamics using Artificial Neural Network, May 2015
10. Mamta M, Lalit MG, Kaur J (2018) Monitoring the impact of economic crisis on crime in india using machine learning. *Comput Econ* 53(4):1467–1485
11. Kaur J, Mamta M (2019) A new feature selection method based on machine learning technique for air quality dataset. *J Stat Manage Syst* 22(4):697–705

12. Wang X, Liu T, Zheng X, Peng H, Xin J, Zhang B (2018) Short-term prediction of groundwater level using improved random forest regression with a combination of random features. *Appl Water Sci* 8(5):1–12
13. Models T et al (2012) Department of Computer Science and Engineering Indian Institute of Technology Bombay June 2012
14. Sakizadeh M, Klammler H (2019) Trend analysis and spatial prediction of groundwater levels using time series forecasting and a novel spatio-temporal method
15. Nayak PC, Rao YRS, Sudheer KP (2006) Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water Resour Manage* 20(1):77–90
16. Mohammad M, Ghazavi R (2015) A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resour Manage* 29(4):1315–1328
17. Rajesh R, Murthy TRS, Raghavan BR (2005) Time series analysis to monitor and assess water resources: a moving average approach. *Environ Monit Assess* 109(1–3):65–72
18. Neto DC, Chang HK, Genuchten MTV (2016) A mathematical view of water table fluctuations in a shallow aquifer in Brazil. *Groundwater* 54(1):82–91
19. Time Series and Forecasting Time Series • A time series is a sequence of measurements over time, usually obtained Components of a Time Series • Secular Trend, pp 1–24
20. <http://cgwb.gov.in/Ground-Water/Groundwater%20Year%20Book%202017-18.pdf>
21. <http://cgwb.gov.in/Ground-Water/Groundwater%20Year%20Book%202016-1.pdf>
22. [http://mowr.gov.in/sites/default/files/AR\\_CGWB\\_2014-2015\\_1.pdf](http://mowr.gov.in/sites/default/files/AR_CGWB_2014-2015_1.pdf)

# Prediction of Air Quality Index Using Hybrid Machine Learning Algorithm



Jasleen Kaur Sethi and Mamta Mittal

**Abstract** Air pollution is an acute problem which leads to detrimental effects on human health and living conditions. Therefore, there is a need to monitor the pollution levels to inform people about the status of current air quality. This is done by an index called Air Quality Index (AQI) that maps the concentration of various pollutants into single value. To predict the AQI, a hybrid machine learning algorithm has been proposed in this paper in which the cluster classifications computed by  $k$ -means clustering algorithm are used as an input to support vector machines (SVM) algorithm. To perform the experimental work, three-year (January 2016 to January 2019) air quality data of Gurugram (Haryana) has been utilized after preprocessing it by scaling. The obtained results of the hybrid approach have been compared to the traditional SVM algorithm. Based on the empirical study, the hybrid algorithm prediction performance is better than SVM algorithm. It has been observed that the accuracy of proposed algorithm is found to be 91.25% as compared to the SVM algorithm with an accuracy of 65.93%.

**Keywords** Air quality index · Supervised learning · Unsupervised learning ·  $K$ -means clustering · Support vector machines

## 1 Introduction

One of the harmful effects of the development of science and technology in the urban areas is air pollution which affects both human health and plants [1]. The tool used to inform people about the air quality is known as Air Quality Index

---

J. K. Sethi (✉)

University School of Information, Communication & Technology, Guru Gobind Singh

Indraprastha University, New Delhi 110078, India

e-mail: [jasleenkaursethi@gmail.com](mailto:jasleenkaursethi@gmail.com)

M. Mittal

Department of Computer Science & Engineering, G. B. Pant Government Engineering College, New Delhi 110020, India

e-mail: [mittalmamta79@gmail.com](mailto:mittalmamta79@gmail.com)

(AQI). This index is calculated from concentration of number of pollutants [2, 3]. In literature, a lot of research has been focussed on hybrid supervised and unsupervised machine learning algorithms for air quality prediction. Tamas et al. [4] predicted the concentration of many pollutants based on clustering and artificial neural network at Corsica Island in the western Mediterranean Sea. Two hybrid algorithms using multi-layer perceptron (MLP) were proposed, and the results were compared to the model based on traditional MLP. It was found that hybrid models based on hierarchical clustering and MLP outperform traditional MLP for prediction of ozone and particulate matter. Bougoudis et al. [5] proposed a model named as Easy Hybrid Forecasting (EHF) used to predict the levels of air pollutants at Athens based on meteorological and temporal parameters. EHF does not require data from sensors to forecast the pollutants and finds application in smartphones. Kolehmainen et al. [6] proposed a prediction model to perform next day air quality forecast using the pollutants and meteorological parameters data based on self-organizing map, fuzzy clustering and MLP. It was concluded that model based on pollutants is more reliable than that based on particulate matter. Bougoudis et al. [7] predicted the concentration of air pollutants based on the combination of semi-supervised classification and clustering. The proposed hybrid model gave better results when tested on air quality data of Greece in comparison with the traditional techniques. This work aims to predict the AQI of Gurugram, Haryana, using hybrid machine learning algorithm based on  $k$ -means clustering and support vector machine.

The study area and the dataset used in the work have been presented in the following section. Section 3 discusses the  $k$ -means clustering algorithm and support vector machines algorithm. The hybrid methodology for AQI prediction has been presented in Sect. 4. The global results of both the hybrid and traditional model have been discussed in Sect. 5. The last section discusses the conclusion.

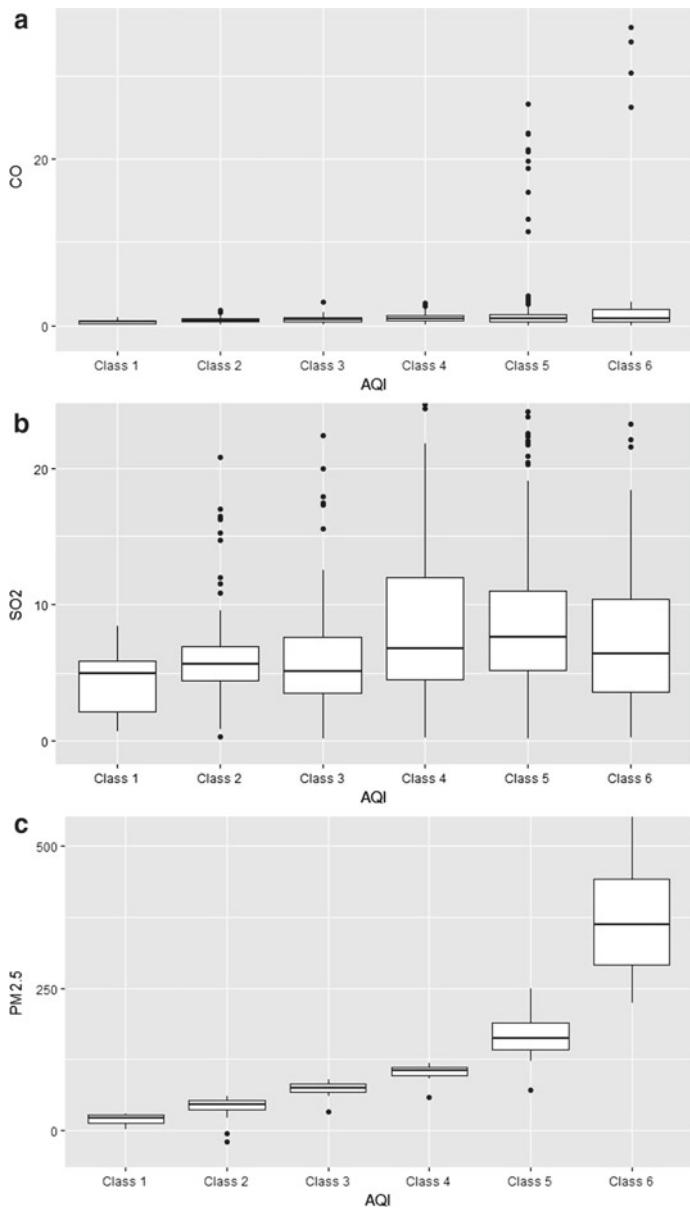
## 2 Study Area and Air Quality Dataset

According to the World Health Organization (WHO), global pollution database, India, has the maximum number of polluted cities in the world [8]. One of the cities in the aforementioned database is Gurgaon (officially called Gurugram) which located at  $28.4595^\circ$  N,  $77.0266^\circ$  E is a city in Haryana with an area of  $1253\text{ km}^2$  and population of 1,514,085 [9]. The map of Gurugram and its districts is shown in Fig. 1 [10]. The status about the current air quality is estimated by a tool called Air Quality Index (AQI). It maps the concentration of many pollutants to a single value [11]. The data of the study area have been taken from the Central Pollution Control Board (CPCB) Website from January 2016 to January 2019 [12]. The parameters of the dataset include the AQI and the concentration of nitrogen dioxide ( $\text{NO}_2$ ), carbon monoxide (CO), ozone, sulphur dioxide ( $\text{SO}_2$ ) and  $\text{PM}_{2.5}$  and meteorological parameters namely wind speed (WS), absolute temperature (AT) and relative humidity (RH). The multiple box plots of each parameter with the AQI are depicted in Fig. 2.

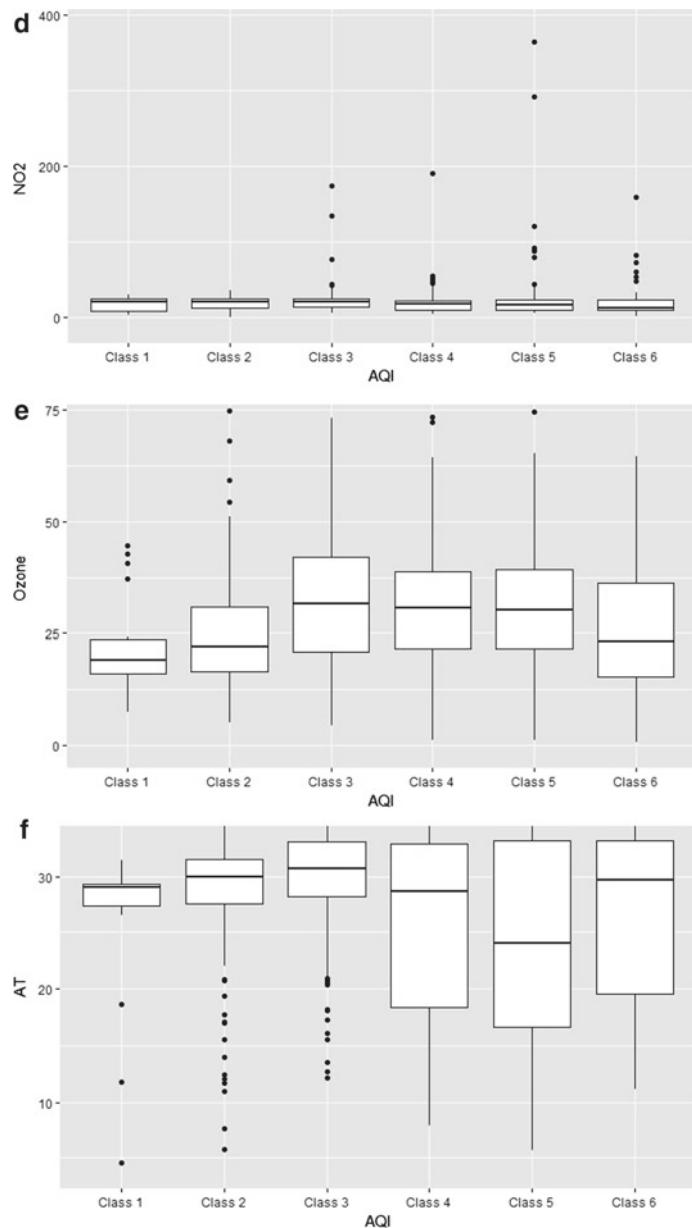


**Fig. 1** Study area map

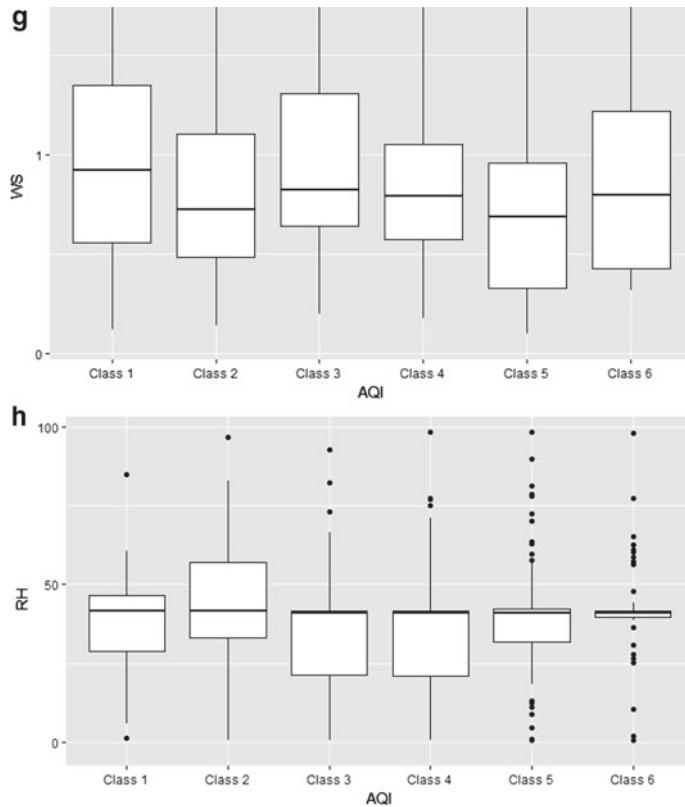
AQI is calculated from the concentration of pollutants namely  $\text{SO}_2$ ,  $\text{NO}_2$ ,  $\text{CO}$ ,  $\text{PM}_{2.5}$  and ozone. The source of  $\text{SO}_2$  includes the natural processes such as volcanic emissions and the combustion of coal and petroleum from industries and automobiles. Sulphur dioxide reacts with other compounds and catalysts to produce acid rain. It has number of health impacts such as wheezing and irritation of throat. Its long-term effects include complications in asthma and heart patients.  $\text{NO}_2$  is a reddish brown gas formed from the combustion of nitrogen containing fuels from vehicles and power plants.  $\text{NO}_2$  also contributes to the formation of haze or smog. It causes problems related to lungs and heart.  $\text{CO}$  is an odourless gas formed by the partial combustion of fossils fuels such as wood or coal. It reduces the capacity of blood to carry oxygen to various parts of the body causing drowsiness, and when the concentration is very high, it causes death. The sources of fine aerosols ( $\text{PM}_{2.5}$ ) include natural ones like smoke from forest fires or combustion of fuels in the vehicles, while the coarse aerosols ( $\text{PM}_{10}$ ) have sources such as dust blown due to wind. Particulate matter reduces the visibility and has number of health impacts such as respiratory disorders in asthma patients. Secondary pollutants are formed from the photochemical reaction of various primary pollutants. These secondary pollutants consisting of solid particles



**Fig. 2** Box plots of parameters of dataset. **a** Carbon monoxide, **b** sulphur dioxide, **c** PM<sub>2.5</sub>, **d** nitrogen dioxide, **e** ozone, **f** absolute temperature, **g** wind speed, **h** relative humidity



**Fig. 2** (continued)



**Fig. 2** (continued)

and liquid drops are called smog whose major component is ozone which leads to problems with lung and eyes [13, 14].

The box plot is a means to display the various distributions of data where the rectangle shows the first to the third quartile and median is specified by the segment inside the rectangle. The line segments above and below the box signify the minimum and maximum value. From the above box plots, it has been observed that the maximum value of CO, NO<sub>2</sub> and PM<sub>2.5</sub> exists for Class 6 of AQI. For the Class 3 and 4 of AQI, the maximum values exist for ozone and SO<sub>2</sub>, respectively.

### 3 Hybrid Machine Learning Algorithms

Supervised learning is learning with a teacher who is well versed with the environment. The response is computed here based on labelled data. In unsupervised learning, no teacher is present and the classes are created from the unlabelled data

[15]. A hybrid algorithm based on support vector machines (supervised) and  $k$ -means clustering (unsupervised) has been proposed in this paper to predict the AQI. These algorithms have been discussed in the next section.

### 3.1 Support Vector Machines (SVM)

SVM utilizes a hyperplane to separate various classes in order to achieve classification [16]. It reduces the generalization error by maximizing the margin and the distance between the instances of the hyperplane. It has been employed to solve many pattern recognition and function estimation problems. In literature, apart from linear classifier, the SVMs based on radial basis function (RBF), polynomials, and multilayer perceptron (MLP) have been employed [17].

Given a dataset with  $\{z_j\}_{j=1}^M$  output pattern and  $\{x_j\}_{j=1}^M$  input pattern, SVM constructs a classifier as

$$z(x) = \text{sign} \left[ \sum_{j=1}^M a_j z_j \psi(x, x_j) + b \right]$$

where  $a_j$  and  $b$  are constants and  $\psi(\cdot)$  is a function that maps the data into higher dimensional space. This function can be linear, polynomial, or RBF SVM [18, 19]. It assumes that

$$z_j [w^T \psi(x_j) + b] \geq 1, \quad j = 1, 2, \dots, M$$

### 3.2 K-Means Clustering

$k$ -means clustering partitions the dataset into clusters such that the Euclidean distance is minimized. This clustering technique finds application in number of areas such as image processing, data compression and data preprocessing in neural networks [20, 21]. Given a set of data points  $Y = (y_1, y_2, \dots, y_n)$  and a set of cluster centres  $Z = (z_1, z_2, \dots, z_c)$ , the cost function is given by:

$$J = \sum_{i=1}^c \sum_{j=1}^n |y_j - z_i|^2$$

The number of cluster centres is chosen randomly, and then, the distance between each input and centroid is computed. The inputs are assigned to each cluster based on the membership matrix represented by

$$u_{ij} = \begin{cases} 1 & \text{if } |y_j - z_i|^2 \leq |y_j - z_k|^2 \forall k \neq i \\ 0 & \text{otherwise} \end{cases}$$

That is data point  $y_i$  lies in cluster  $i$  if  $z_i$  is the closest centroid amongst all centroids. Then, the cost function according to above equation is computed [22, 23]. The centroids are then updated and the above steps are repeated [24, 25].

## 4 Proposed Hybrid Methodology for AQI Prediction

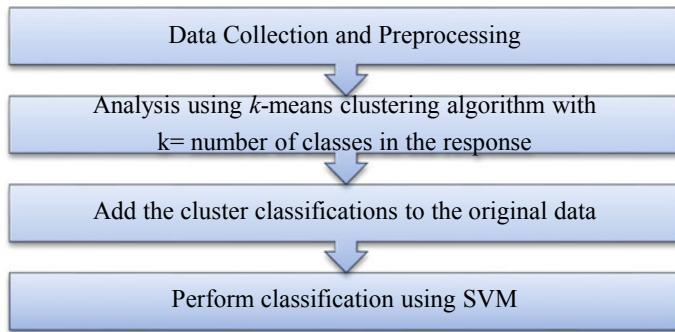
One of the drawbacks of  $k$ -means clustering algorithm is the problem to choose the value of the number of clusters. To overcome this, a hybrid machine learning algorithm has been proposed to predict the AQI where the number of clusters is set to the number of AQI categories. After clustering, the resultant classification is applied to the SVM algorithm. The steps of the proposed hybrid approach are summarized as follows:

- The air quality dataset of Gurugram, Haryana, has been collected from January 2016 to January 2019. From the concentration of different pollutants, the AQI has been computed based on the subindex of each pollutant and then applying the max operator formula. Then, scaling is done to preprocess the data where the values of all parameters are normalized between 0 and 1.
- After normalizing the data,  $k$ -means clustering is applied to find the clusters in the unlabelled data. AQI is divided into six categories namely good, satisfactory, moderately polluted, poor, very poor and severe. Therefore, the number of clusters in  $k$ -means clustering is set to six. Next, based on the results of clustering, the cluster classifications are added to each row to the original data.
- Next, the original AQI parameter is replaced with the cluster classification found. This new dataset is then input to the classification algorithm. Next, the classification is performed using SVM algorithm. These steps have been summarized in Fig. 3.

The performance of the hybrid algorithm is measured based on many performance metrics. The results of the hybrid algorithm have been compared with the traditional SVM algorithm. These results have been discussed in the next section.

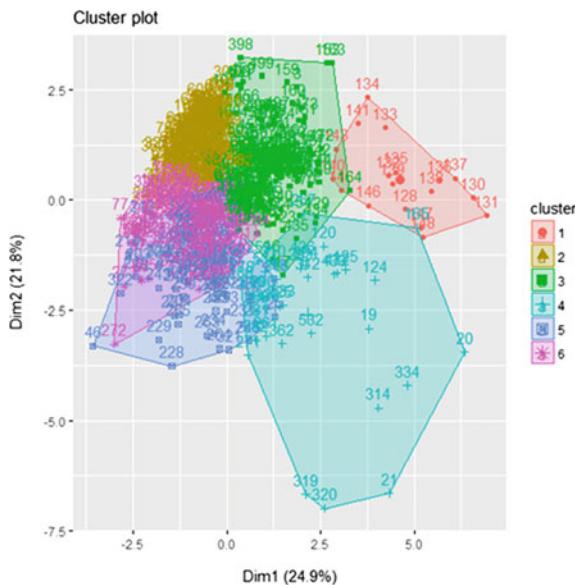
## 5 Results and Discussion

$k$ -means clustering has been applied to the air quality dataset with number of clusters equal to six as there are six categories of AQI. The cluster plot for  $k$ -means clustering is depicted in Fig. 4.



**Fig. 3** Proposed hybrid methodology for AQI prediction

**Fig. 4** Cluster plot for  $k$ -means clustering



A number of metrics like precision, recall, accuracy, error rate and F1 score have been used to evaluate the hybrid algorithm on the dataset of Gurugram. These results have been compared to traditional SVM algorithm. The results of precision and recall for each of the class label for each approach are shown in Table 1. From these results, it has been inferred that proposed algorithm has higher values of precision and recall than the traditional SVM algorithm. Next, the accuracy, error rate and F1 score have been calculated for proposed algorithm. To calculate the F1 score for each technique, the average precision and average recall for all classes has been taken into consideration. These observations are depicted in Table 2.

**Table 1** Precision and Recall values for hybrid algorithm

Class	Precision	Recall	Precision	Recall
	SVM	Hybrid	SVM	Hybrid
1	0	1	0	0.8181
2	0.5468	0.9	0.7446	0.9729
3	0.4745	0.9468	0.5	0.9673
4	0.5106	1	0.4363	0.8421
5	0.7952	0.9047	0.8782	0.8444
6	1	0.8607	0.6764	0.8607

**Table 2** Performance evaluation of proposed hybrid algorithm

Model name	Accuracy	Error rate	F1 score
SVM	0.6593	0.3407	0.656
Hybrid algo ( $k$ -means + SVM)	0.9125	0.0875	0.9089

From the above table, it has been observed that the proposed hybrid algorithm has the higher value of accuracy and F1 score compared to SVM. Further, this algorithm has lower value of error rate.

## 6 Conclusion

Air pollution is a major concern for human health and the environment. Therefore, the pollution levels should be continuously monitored using Air Quality Index. To predict this index, a hybrid approach which is based on  $k$ -means clustering and SVM algorithm has been proposed in this study. For the experimental work, the data of Gurugram from CPCB have been taken. This dataset has also been preprocessed by scaling all values between 0 and 1. It has been found that the proposed hybrid algorithm outperformed the traditional SVM algorithm. The proposed algorithm has an accuracy of 91.25%, an error rate of 0.0875, and F1 score of 0.9089 when compared to SVM with an accuracy, error rate and F1 score of 65.93%, 0.3407 and 0.656, respectively, on same dataset.

## References

1. Wang D, Wei S, Luo H, Yue C, Grunder O (2017) A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. *Sci Total Environ* 580:719–733

2. Singh KP, Gupta S, Rai P (2013) Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmos Environ* 80:426–437
3. Streets DG, Fu JS, Jang CJ, Hao J, He K, Tang X et al (2007) Air quality during the 2008 Beijing Olympic games. *Atmos Environ* 41(3):480–492
4. Tamas W, Nottou G, Paoli C, Nivet ML, Voyant C (2016) Hybridization of air quality forecasting models using machine learning and clustering: an original approach to detect pollutant peaks. *Aerosol Air Qual Res* 16:405–416
5. Bougoudis I, Demertzis K, Iliadis L (2016) Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning. *Integr Comput-Aided Eng* 23(2):115–127
6. Kolehmainen M, Martikainen H, Hiltunen T, Ruuskanen J (2000) Forecasting air quality parameters using hybrid neural network modelling. *Environ Monit Assess* 65(1–2):277–286
7. Bougoudis I, Demertzis K, Iliadis L, Anezakis VD, Papaleonidas A (2016) Semi-supervised hybrid modeling of atmospheric pollution in urban centers. In: International conference on engineering applications of neural networks, Sept 2016. Springer, Cham, pp 51–63
8. <https://timesofindia.indiatimes.com/city/delhi/14-of-worlds-15-most-polluted-cities-in-india/articleshow/63993356.cms>. Accessed on 10 Aug 2019
9. <https://gurugram.gov.in/>. Accessed on 10 Aug 2019
10. <https://haryanamap.wordpress.com/tag/map-of-haryana/>. Accessed on 10 Aug 2019
11. Sethi JK, Mittal M (2019) A new feature selection method based on machine learning technique for air quality dataset. *J Stat Manage Syst* 22(4):697–705
12. Central Pollution Control Board (CPCB), Government of India. <http://cpcb.nic.in/>. Accessed on 10 Aug 2019
13. Chen Z, Cai J, Gao B, Xu B, Dai S, He B, Xie X (2017) Detecting the causality influence of individual meteorological factors on local PM<sub>2.5</sub> concentration in the Jing-Jin-Ji Region
14. Hu K, Rahman A, Bhrugubanda H, Sivaraman V (2017) HazeEst: machine learning based metropolitan air pollution estimation from fixed and mobile sensors. *IEEE Sens J* 17(11):3517–3525
15. Mittal M, Goyal LM, Sethi JK, Hemanth DJ (2019) Monitoring the impact of economic crisis on crime in India using machine learning. *Comput Econ* 53(4):1467–1485
16. Sethi JK, Mittal M (2019) Ambient air quality estimation using supervised learning techniques. SIS, EAI
17. Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. *Neural Process Lett* 9(3):293–300
18. Lu W, Wang W, Leung AY, Lo SM, Yuen RK, Xu Z, Fan H (2002) Air pollutant parameter forecasting using support vector machines. In: Proceedings of the 2002 international joint conference on neural networks. IJCNN'02 (Cat. No. 02CH37290), May 2002, vol 1. IEEE, pp 630–635
19. Lu WZ, Wang WJ (2005) Potential assessment of the “support vector machine” method in forecasting ambient air pollutant trends. *Chemosphere* 59(5):693–701
20. Mittal M, Goyal LM, Hemanth DJ, Sethi JK (2019) Clustering approaches for high-dimensional databases: a review. *WIREs Data Mining Knowl Discov*
21. Goyal LM, Mittal M, Sethi JK (2016) Fuzzy model generation using subtractive and fuzzy C-Means clustering. *CSI Trans ICT* 4(2–4):129–133
22. Mittal M, Sharma RK, Singh VP, Goyal LM (2017) Modified single pass clustering algorithm based on median as a threshold similarity value. In: Collaborative filtering using data mining and analysis. IGI Global, pp 24–48
23. Mittal M, Sharma RK, Singh VP (2015) Modified single pass clustering with variable threshold approach. *Int J Innov Comput Inf Control* 11(1):375–386
24. Mittal M, Sharma RK, Singh VP, Kumar R (2019) Adaptive threshold based clustering: a deterministic partitioning approach. *Int J Inf Syst Model Design (IJISMD)* 10(1):42–59
25. Mittal M, Sharma RK, Singh VP (2019) Performance evaluation of threshold-based and k-means clustering algorithms using iris dataset. *Recent Patents Eng* 13(2):131–135

# Employing Blockchain in Rice Supply Chain Management



M. Vinod Kumar, N. Ch. Sriman Narayana Iyengar, and Vishal Goar

**Abstract** Blockchain is considered to be the next paradigm in information technology after mainframes, computer, Internet, and smartphones. With so much hype around cryptocurrencies the mechanism on which these were built, “The Blockchain” has drawn attention of many scientists and developers around the world. With time, other than cryptocurrency, Blockchain had impact on several other industries such as logistics, health care, real estate, legal industries, etc. One such application leads to employing Blockchain technology into food supply chain. Rice is the major food consumed by people in India. Rice supply chains play a vital role in supplying rice from manufacturer to consumers. Hence, for attaining a corruption-free, transparent, and efficient rice supply chain, Blockchain is to be employed in the functioning of these supply chains such that safety of rice can be monitored at different stages involved in the supply chain. In this paper, a theoretical study is shown on how Blockchain can be integrated into regular rice supply chain.

**Keywords** Crypto currencies · Blockchain · Rice supply chain

## 1 Introduction

In 2008, a person known by his pseudonym Satoshi Nakamoto has created Bitcoin a digital currency which is intended to make money transfers directly from one person to other person without any intermediaries like banks or financial institutions to be present in-between. Blockchain is the technology behind the Bitcoin which

---

M. Vinod Kumar · N. Ch. Sriman Narayana Iyengar (✉)

Department of Information Technology, Sreenidhi Institute of Science and Technology,  
Yamnampet, Ghatkesar, Hyderabad, Telangana, India

e-mail: [srimannarayananach@sreenidhi.edu.in](mailto:srimannarayananach@sreenidhi.edu.in)

M. Vinod Kumar

e-mail: [mvk2407@gmail.com](mailto:mvk2407@gmail.com)

V. Goar

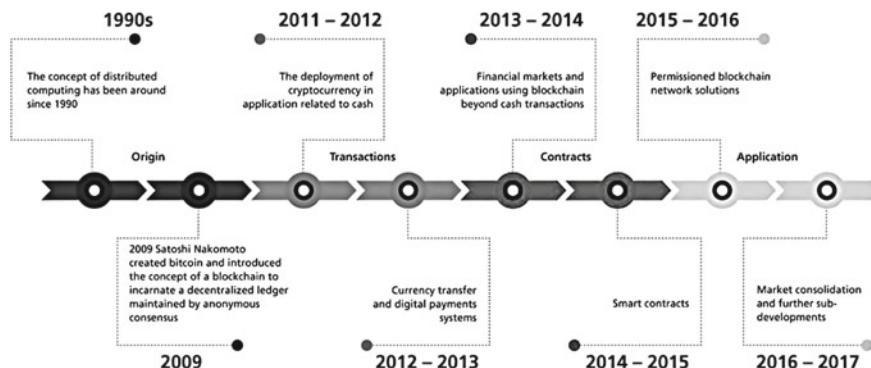
Department of CA, Engineering College, Bikaner, Rajasthan, India

e-mail: [dr.vishalgoar@gmail.com](mailto:dr.vishalgoar@gmail.com)

gives it ability to perform electronic peer-to-peer cash transactions and maintain ledgers in decentralized and distributed manner such that the ledgers are accessible to everyone while keeping them secure. Blockchain in simple words is a distributed network of databases which is not governed by any central authority and makes the data immutable once linked with the Blockchain by not allowing anyone to tamper it. There are different types of Blockchain based on dependencies such as public, private, and consortium.

Ethereum Blockchain is second-generation Blockchain technology developed by Vitalik Buterin in 2014. Ethereum gives an ability to develop DAPP (decentralized application) and execute smart contracts over EVM (Ethereum Virtual Machine). Smart contract is a type of digital agreement running on top of Ethereum Blockchain containing a set of rules for making fare trades between different parties. As Bitcoin, Ethereum too has its own cryptocurrency called “Ether” which also performs peer-to-peer transactions. This robust technology can revolutionize the mode of doing businesses if integrated in the functioning of several industries such as logistic, pharmaceutical, election, real estate, politics, etc. Hence, in this paper, an attempt is made to adopt Blockchain into rice supply chain management system. As rice is majorly consumed as food in India, it is responsibility of rice supply chains to deliver a quality product to consumers. The lack of transparency in the processes and mutual cooperation between the several entities involved in the rice supply chain is making it inefficient, thus people leaving hopes in the markets. Integrating Blockchain with the existing rice supply chain creates a transparency in the network of flow of rice from farmers to consumer, and immutable track of records is created such that any person belonging to network with an Internet connection can audit and validate the quality of the product when received (Fig. 1).

## BLOCKCHAIN HISTORY

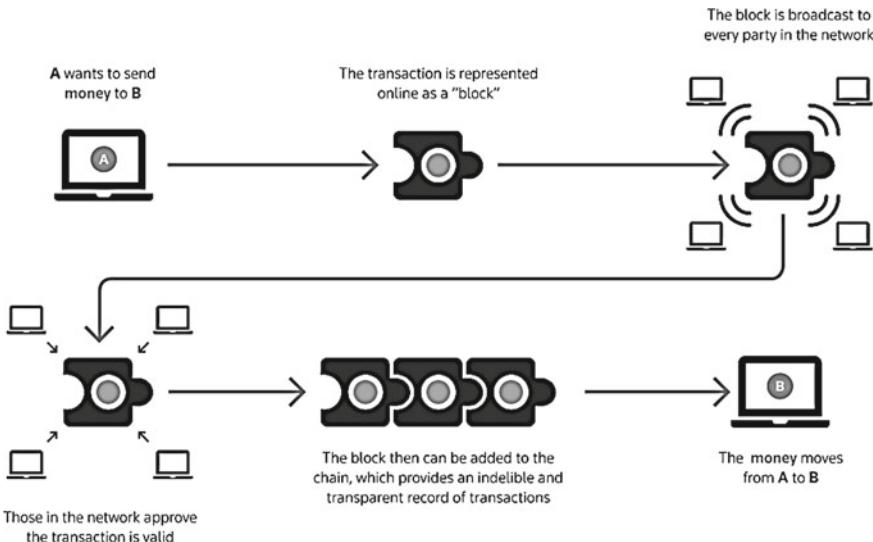


**Fig. 1** A history of blockchain technology. *Source* Accenture

## 2 Literature Survey

A person with pseudonym Satoshi Nakamoto has created Bitcoin, a fully distributed digital currency which does not belongs to any central authority and also does not require any intermediaries such as banks and financial institutions to carry out transactions [1]. It is a peer-to-peer electronic cash system that allows a person to directly transfer his money to other person without any third party to be present in between. With outbreak of cryptocurrencies, the technology on which it were built which is Blockchain has come to lime light. Blockchain is a decentralized and distributed ledger which cannot be tampered and anonymity of person is maintained due to usage of cryptographic hashes [2]. In simple words, Blockchain is a distributed database which is distributed among its peers in the network and synced via the internet. There are different types of Blockchain based on the requirement and type of permissions allotted to participants in a network like public, private, and consortium [3]. Vitalik Buterin through his white paper introduced the 2.0 version of Blockchain to the world. It is similar to that of Bitcoin Blockchain [4]. It is completely decentralized and distributed, and addition to this provides ability to develop decentralized applications DAPPS and write smart contracts. Smart contract is type of digital agreement and autonomous code which executes when a specific condition is met [5]. These smart contracts are powered by EVM (Ethereum Virtual Machine) and written in Solidity. Solidity is the programming language used to develop DAPPS and write smart contract.

Like Bitcoin, Ethereum has its own cryptocurrency called “Ether”. Knowing the potential of Blockchain businesses of all kinds are getting creative. Due to its robustness and unique features, several industries like pharma, trading, and logistics are integrating Blockchain into their business practices [6]. Blockchain technology can even improve functioning of construction corporations by using three types of Blockchain applications which are notarization-related, provenance-related and transaction-related [7]. The Blockchain technology can be embodied in Customs Department such that all participants involved in the system have a single record to check the authenticity and validate the product's integrity toward cross-border transfers [8]. As present situation of logistics and supply chain industries can be described as challenging, employing Blockchain into their operations can increase chances of ease-of-doing business and fight frauds in the system [9]. Blockchain can also be used in food supply chains where data related to product is kept secured and immutable. A researcher in his paper used Blockchain technology to create transparency in the supply chain network by building a conceptual model using the Unified Theory of Acceptance and Use of Technology (UTAUT) and developed traceability system to track the product in supply chain [10]. Rice is a basic grain consumed as food in India. Rice supply chains play an important role in bringing rice from farmers to consumers. Intermediaries such as middle agents, rice processing companies, distributors, and retailers are involved in rice supply chain [11]. Going through different stages and maintaining product's integrity throughout the different processes involved in supply chain are the main aim of rice supply chain. Adapting Blockchain into its functioning



**Fig. 2** Illustration of a blockchain transaction. *Source* Financial Times

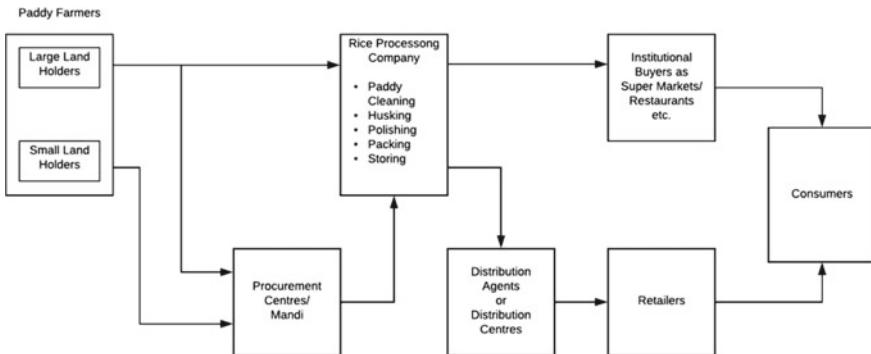
may help supply chains fight frauds and limit the inefficiency arising in supply chains [12] (Fig. 2).

### 3 Traditional Rice Supply Chain Model

The participants involved in rice supply chain are farmers, mandi or middle agents, RPC (rice processing company), distributors, retailers, and consumers. Firstly, farmers cultivate paddy and sell the harvested paddy at local mandis or various procurement centers. Mandi people collect the paddy and supply it to RPC. At RPCs, paddy gets processed and converted to rice. After that many distributors approach RPC and buy rice in huge quantity. From distributors, rice is being supplied to local retailers, and from retailers, rice reaches to public. In case of industrial buyers or restaurants, they directly buy rice from RPCs in huge quantity. Figure 3 explains in detail the journey of rice from farmer to consumer passing through several intermediaries.

### 4 Problems with Existing Model

- The main problem in the traditional rice supply chain is existence of several intermediaries in the process of supplying rice from manufacturer to consumer.



**Fig. 3** Traditional rice supply chain

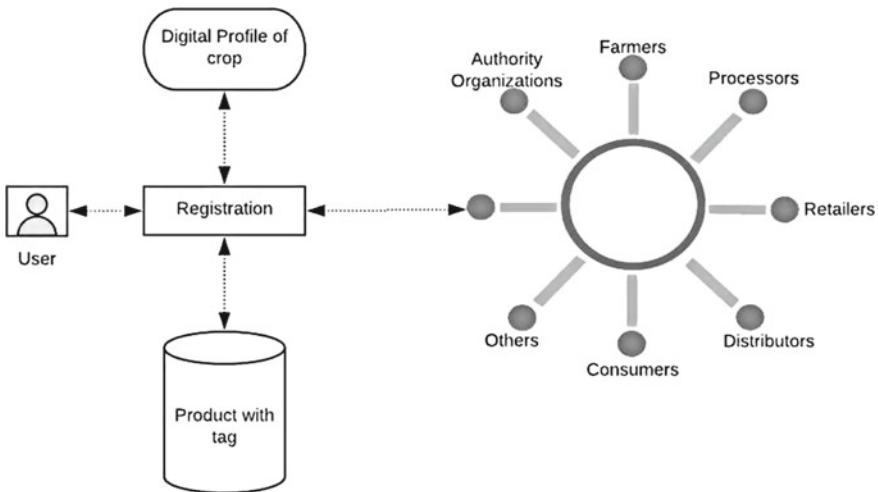
- No proper coordination and mutual cooperation between the participants involved in the supply chain system leads to rise of many issues.
- The opaque nature of entire processes involved in the supply chain of rice is making consumer lose trust on supply chains.
- Greater complexity in the functioning of different processes is main cause for inefficiency in existing rice supply chain.

## 5 Working Model

All the participants involved in the supply chain management system are to be pre-registered in the Blockchain with a unique ID such that they have feasibility to upload data related to product (rice) on to the Blockchain. Next when the paddy is being harvested and packed into bags, they are labeled with specific tags. With the same Tag ID named, a digital profile of the product is created on Blockchain by farmer at initial stage and all the data related to product are uploaded on to that digital profile. Stages involved in supplying rice from farmer to consumer are procuring, processing, distributing, and retailing. At each stage, data related to product are continuously updated on the Blockchain at specific intervals such that a traceability system is built to track the product in supply chain (Fig. 4).

### Manufacturing

At this stage, farmer creates digital profile of the product on Blockchain and uploads data related to paddy like type of seeds used, planting time, fertilizers used, and plucking time. Tags are being inserted on the bags of paddy and made to sync with their related digital profiles on blockchain. Later farmers sell those bags at local mandi or procuring centers, and with the help of “Smart Contract”, ownership is exchanged, and those data are uploaded on the Blockchain.



**Fig. 4** Registration process

### Procuring

After receiving bags from farmers, the details regarding logistics, warehousing, location, etc., are being continuously updated on to the digital profile of the product at specific intervals.

### Processing

At processing center which are called as rice processing company or rice mill, paddy is collected and being processed. It will go through multiple stages such as cleaning, husking, and polishing to extract rice from paddy. All the data related to different processes in converting paddy to rice are being updated to digital profile of the product which is stored on the Blockchain. All the processes involved are shown in Figure 2. Further, the processed rice is packed in bags and new tags are inserted on to the bags and are made that new tags get synced with previous tags and get configured such that the further details are entered using this new tags. After packing rice into bags, these bags are supplied to distributors.

### Distributing

Distributor purchase rice from RPC and information related to logistics, warehousing, quality of rice is continuously being synced with the digital profile of the product and product status can be tracked and unusual activity can be recorded.

### Retailing

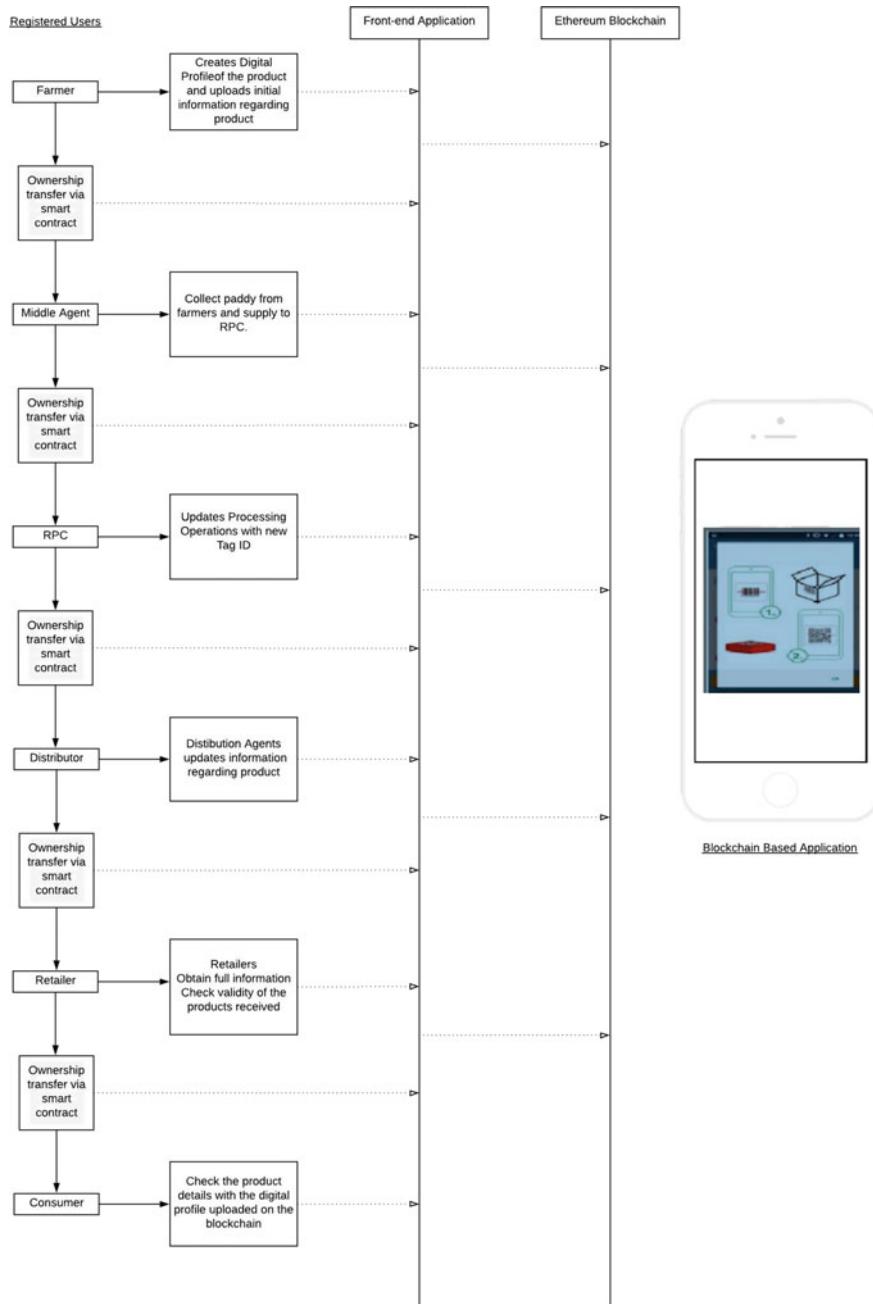
When retailers receive the rice bags, they nearly obtain all the information related to specific rice and can validate the activities involved in rice supply chain and assure that the validity of product to consumer. As entire supply chain is transparent, a person who acts as a node to Ethereum Blockchain can access all the information

related to specific rice bags. Integrating blockchain into rice supply chain is given in Fig. 5 and attributes are given in Fig. 6.

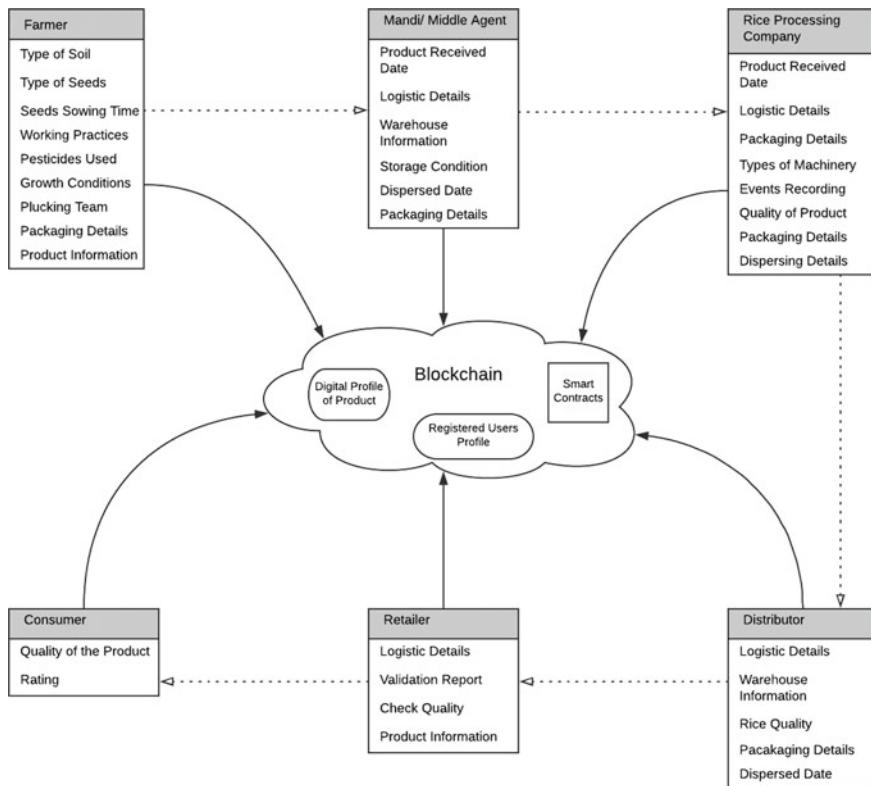
## 6 Technical Aspect

As the proposed model to implement, firstly we need to build a private Ethereum Blockchain. Geth (Go-ethereum) can be used to build the private blockchain. Geth also helps in creating nodes, performs operations such as mining, and is also used to create Ethereum clients to connect to public and test networks. Geth is configured to run locally without connecting to any network on the Internet. Every blockchain has a genesis block or the first block. This block does not have any parent and is the first block of the chain. A “genesis.json” file is required to create the first block.

## **Code for creating genesis block**



**Fig. 5** Integrating blockchain into rice supply chain



**Fig. 6** Attributes uploaded by the participants to the blockchain

```

“gasLimit”: “0xffffffff”,
“alloc”: {
}
  
```

**The other components required to function blockchain are:**

- Solidity (Ethereum)
- Metamask (Ethereum wallet)
- Rinkeby test network (use rinkeby faucet to get ethers on rinkeby network)
- Infura
- Remix—Solidity IDE
- IPFS
- Web3JS

```

1 pragma solidity >0.4.23;
2
3 contract RiceSupplyChain
4 {
5
6     event PerformCultivation(address indexed user, address indexed batchNo);
7     event DoneInspection(address indexed user, address indexed batchNo);
8     event DoneHarvesting(address indexed user, address indexed batchNo);
9     event DoneImporting(address indexed user, address indexed batchNo);
10    event DoneProcessing(address indexed user, address indexed batchNo);
11
12
13    modifier isValidPerformer(address batchNo, string role) {
14        require(keccak256(supplyChainStorage.getRole(msg.sender)) == keccak256(role));
15        require(keccak256(supplyChainStorage.getNextAction(batchNo)) == keccak256(role));
16        ...
17    }
18
19
20    SupplyChainStorage supplyChainStorage;
21
22    constructor(address _supplyChainAddress) public {
23        supplyChainStorage = SupplyChainStorage(_supplyChainAddress);
24    }
25
26
27    function getNextAction(address _batchNo) public view returns(string action)
28    {
29        (action) = supplyChainStorage.getNextAction(_batchNo);
30    }
31
32
33 }
```

Remix IDE Browser for executing Smart Contracts.

## Benefits of Integrating Blockchain into Rice Supply Chain

- A mutual cooperation is established between the entities involved in the supply chain system.
- A transparent network is created such that any concerned people will audit and get access to the information regarding rice supply chain.
- A traceability system is built such that product can be tracked throughout the different processes involved in the rice supply chain.
- By using smart contracts, fare trades are initiated between the people involved in the supply chain.
- By integrating Blockchain technology into rice supply chain, we create an environment where there will be transparency in the network and safety of the product is regularly monitored.

## 7 Conclusion

In this paper, a theoretical study is shown on how Blockchain technology can be integrated with the rice supply chain management. As rice is a basic grain served as food in India, it is responsibility of rice supply chains to deliver a product of premium quality to public. As blockchain is decentralized and no one governs it, tampering the data related to flow of product in rice supply chain is impossible. Using Ethereum Blockchain technology in rice supply chain management, flow of rice can be regularly monitored and a transparency is created in the network such that chances of occurring frauds in supply chain are reduced to maximum. Blockchain is a revolutionary technology which challenges the traditional way of doing businesses and helps in improving the functioning of various industries. The main objective of this paper is to make familiar with the potential of blockchain and to know what changes it would

bring to food industries if applied into practice. Integrating blockchain technology in food industry will promise of delivering a quality product to public.

## References

1. Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>, white paper
2. Zyskind G, Nathan O (2015) Decentralizing privacy: using blockchain to protect personal data. Security and privacy workshops (SPW), 2015 IEEE, CA, USA, May 21–22, pp 180–184
3. Types of blockchain, <https://blog.darwinlabs.io/types-of-blockchain-public-private-and-permissioned-5b14fbfe38d4>
4. Buterin V (2014) A next-generation smart contract and decentralized application platform. <https://github.com/ethereum/wiki/wiki/White-Paper>, white paper
5. Christidis K, Devetsikiotis M (2016) Blockchains and smart contracts for the internet of things. IEEE Access
6. Bocek T, Rodrigues BB (2017) Blockchains everywhere—a use-case of blockchains in the pharma supply-chain. In: 2017 IFIP/IEEE symposium on integrated network and service management (IM2017), Portugal, 8–12 May 2017, pp 772–777
7. Wang Jun, Peng Wu et al (2017) The outlook of blockchain technology for construction engineering management. Front Eng Manage 4(1):67–75
8. Okazaki Y (2018) Unveiling the potential of blockchain for customs. WCO Research Paper No. 45, June 2018
9. Hackius N, Petersen M (2017) Blockchain in logistics and supply chain: trick or treat? In: Hamburg international conference of logistics (HICL), 23 Oct 2017
10. Francisco K, Swanson D (2018) The supply chain has no clothes: technology adoption of blockchain for supply chain transparency, [www.mdpi.com/journal/logistics](http://www.mdpi.com/journal/logistics), Jan 2018
11. Somashekhar IC, Raju JK (2014) Agriculture supply chain management: a scenario in India. Res J Soc Sci Manage RJSSM 04(07):89–99
12. Tse D, Zhang B et al (2017) Blockchain application in food supply information security. In: 2017 IEEE international conference on industrial engineering and engineering management (IEEM), 10–13 Dec 2017

# Supervised Learning Method and Neural Network Algorithm for the Analysis of Diabetic Mellius and its Comparative Analysis



J. Jayashree, J. Vijayashree, N. Ch. Sriman Narayana Iyengar, and Vishal Goar

**Abstract** Diabetes is the key critical issue needs to be concerned for various problems in our body. Increase in glucose and fructose content in our body results in diabetes mellitus. When a body generates higher insulin level than the required, it results in increased urination and excessive thirstiness which in turn results in kidney failure and other cardio-related issues. Many research agencies invested their funds on defining the predictive methodology and finding the root cause of those results in mellitus. Mellitus results in the highest mortality rate compared to any other disease reported by the health organizations across the globe. In this, the predictive methodologies, various classification techniques are discussed, and the results are analyzed. The classification methodology could be on medications, food habits, personal behaviors, age factors and so on. The datasets are processed and analyzed with the neural network algorithms, and the results are compared with one another. The datasets are taken from the National Family Health Survey results published during the period of 2016–2017. The result implies that men between ages 15–49 among 1 billion people have reported with diabetes mellitus. Diagnose and forecast on this disease are done by recognizing the pattern formation and grouping the similar structures. Various algorithmic techniques like M-layer perceptron, nearest neighbor, vector machines, data regressions, binary regression and their accuracy of forecast, speed and sensitivity are calculated, analyzed and compared to define the accurate prediction methodology over a short span of time. The forecast methodologies are focussed

---

J. Jayashree · J. Vijayashree

School of Computer Science and Engineering, VIT, Vellore, Tamil Nadu, India

e-mail: [jayashree.j@vit.ac.in](mailto:jayashree.j@vit.ac.in)

J. Vijayashree

e-mail: [vijayashree.j@vit.ac.in](mailto:vijayashree.j@vit.ac.in)

N. Ch. Sriman Narayana Iyengar (✉)

Department of Information Technology, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India

e-mail: [srimannarayananach@sreenidhi.edu.in](mailto:srimannarayananach@sreenidhi.edu.in)

V. Goar

Department of CA, Engineering College Bikaner, Bikaner, Rajasthan, India

e-mail: [dr.vishalgoar@gmail.com](mailto:dr.vishalgoar@gmail.com)

to provide solutions to avoid the intensive care system provided proper medications with a long duration when it is been predicted to be a risk factor. A statistical method of analyzing is performed for the comparative analysis. The learning and training methodologies are discussed in this system. Accuracy, specificity, sensitivity are the key parameters to define the best forecast methodology. Classification on association, regression techniques and neural algorithmic techniques is analyzed and compared to refine the best predictive forecast methodology by processing 30 samples across the states of India with focus on determining the type of mellitus along with the accuracy on definition. The forecast data utilized to define the type of mellitus and the prediction on critical measures over a period of time.

**Keywords** Mellitus · Neural algorithms · Mellitus classification · M-layer perceptron · Regression techniques · Nearest neighbor · Learning techniques · MATLAB

## 1 Introduction

The diabetes mellitus is a serious health issue needs to be appropriately diagnosed and treated with proper medications and regular checkup. Mellitus is caused due to the excessive insulin secretion or due to the lack of insulin needed for the body biological balance. The excessive insulin secretion is due to improper physical fitness, increased intake of starch contents, excessive carbohydrates and low secretion is due to excessive dilution, enormous medication intake, skipping the meals, less food intake than required level. The normal level of glucose for all the conditions of non-diabetic and diabetic conditions is given in Table 1. Exceeding these level results in various clinical disorders in the human body. The issue could be from normal to extreme intensive conditions. The intensive conditions could be damage of cells in kidney, nervous disorder and defective retinal aperture. The severity can be classified with three types of disorders. The level-1 disorder results in improper insulin secretion and defects pancreas in generating the insulin. This type disorder is caused by the genetic system and causes damage in the nerve cells. The nerve cell adhered back to the retina is defected by the damage in nerve cell and results in blindness. Due to improper secretion of insulin level, pancreas affected which in turn affects the kidney functionality [1].

**Table 1** Normal glucose level recommended by the national clinical Institute

Target condition	Upon day start (mmol/L)	Pre-meal (mmol/L)	Post-meal (mmol/L)
Non-diabetic		4–5.9	<7.8
Type 2 diabetic		4–7	<8.5
Type 1 diabetic	5–7	4–7	5–9
<12 years with type 1 diabetes	4–7	4–7	5–9

Level 2 mellitus disorder usually occurs in the teenage is due to the excessive weight, hypertension, fat contents and improper BMI. Combination of this level of mellitus disorder along with hypertension results in cardiac vascular disorders such as cardiac arrest and myocardial infarctions. The insulin secretion at the pancreas is not sufficient for the normal body functionality. Also, the excessive weight imposes the improper secretion of insulin level. The improper secretion is by the fat muscles and liver which is formed due to the improper food intake and restless sleep. However, this level of mellitus disorder can be controlled by implementing proper diet, physical fitness, regular and timeliness food intake and sleep. There is another type of mellitus which only causes during the pregnancy is called gestational mellitus. This mellitus disorder results in the inhabitant growth of the child inside the womb.

Gestational mellitus should be brought under the control and a high intensive care, and physical fitness and precautionary medication without excessive medication intake should be followed. There is a 55–60% chance for the continuation of gestational mellitus into level 2 mellitus disorder during the postpregnancy. In the absence of intensive care on gestational mellitus increases the risk factor of unborn baby which in turn results in excessive weight gain at the womb stage, breathing issue, level 2 mellitus state of mellitus once after born. The gestational mellitus also increases the risk factor for the normal delivery of baby.

Various research works on diabetes prediction are represented with important terminologies in section II. The neural algorithms and training procedures are represented in section III. The section III has various algorithmic techniques which are implemented and processed with the diabetic datasets on various classifications with 14 instances as datasets.

## 2 Literature Review

In this section, various methodologies involved in the prediction of mellitus disorder using various neural algorithmic techniques are analyzed. This results in enhanced usage of various datasets for the different neural algorithmic techniques to improvise the system on accuracy, specificity and sensitivity.

Classification of datasets with 100 observations and seven class instances like pressure level, serum albumin and age, type of diabetic disorders, tri-fold thickness, cholesterol, BMI and personal behavior are taken into consideration. Binary logistic regression methodology, KNN, MLP are the neural algorithmic techniques taken for the comparative analysis. The results found to be 0.809 sensitive, 0.56 specific and 0.69 accurate for binary logistic regression methodology and 0.825 for both MLP and KNN. The results found to be more accurate in KNN compared to MLP with 72–81 instances and 19–28 incorrect or noise data [2].

Naive Bayes method of forecasting methodology is analyzed with 760 samples using the PIMA Indian level II mellitus disorder datasets. Among the 760 samples, only 76 samples were used for the testing purpose and the rest is for the training

the datasets. The accuracy on training the datasets and testing it found to be 89% of training accuracy and 81% of testing accuracy, respectively [3].

### 3 Proposed Work

#### 3.1 Neural Algorithmic Technique for Forecasting Mellitus Disorder with Level State

Multidimensional analysis datasets have been calculated with the help of datasets provided by the National Family Health Survey Committee. There are about ten attributes which are taken into consideration for the forecasting method. This forecast method is accurate to find the type of mellitus occurred with the computed data, generated from the patient with his normal blood test. Various observations on the test have been carried out, and multiple datasets are generated from multiple patients. The datasets and the class instances are described in Sect. 3.1 with the Table 2, attributes and their values.

**Table 2** Attributes and parameters

Attribute No.	Attribute	Parameter type
1	Age	Age parameter
2	BMI	Physical parameter
3	Weight	Physical parameter
4	Glucose	Glycemic parameter
5	Ketone	Cell parameter
6	Peptic acid	Beta cell function
7	Oleic acid	Insulin resistance
8	HbA1	Glycemic parameter
9	Fructosamine	Glycemic parameter
10	Linoleoyl	Insulin resistance
11	Pro insulin	Beta cell parameter
12	Insulin post-meal	Glycemic parameter
13	Insulin pre-meal	Glycemic parameter
14	Hemoglobin	Blood level parameter
15	Glycation gap	Glycemic parameter

**Table 3** Risk values and their parameters

Value	Indication Level
$<3.9 \text{ mmol/L}$	Low fasting insulin level
$3.9 < x < 6 \text{ mmol/L}$	Normal fasting level for adults
$6.1 < x < 6.9 \text{ mmol}$	Pre-diabetic fasting level
$>7.0 \text{ mmol/L}$	Presence of diabetic mellitus
$3.9 < x < 7.8 \text{ mmol/L}$	Post-meal normal level
$7.9 < x < 10.9 \text{ mmol/L}$	Border post-meal level
$<3.9 \text{ mmol/L}$	Hypoglycemia initial state
$2.8 \text{ mmol/L}$	Hypoglycemia fasting level
$<2.8 \text{ mmol/L}$	High risk of insulin
$8 < x < 11 \text{ mmol/L}$	Determines early diabetic mellitus
$>11 \text{ mmol/L}$	Determines generalized diabetic mellitus

### 3.2 Input Datasets

In general, dataset instances are age, blood group, hemoglobin level, glucose/insulin level, weight, BMI, Ketone level, serum albumin, post-meal and pre-meal test insulin level, blood pressure level, HbA1, Fructosamine, fatty acid, oleic acid, Linoleoyl-GPC, peptides, proinsulin which are beta cells function estimations of insulin level from the blood test analysis [4].

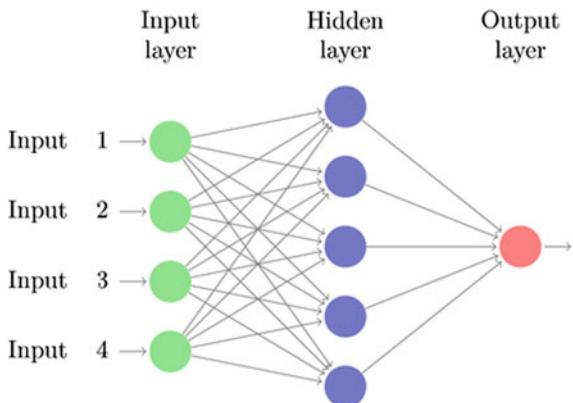
The laboratory test results of patients and the risk factors on different parameters analyzed from the blood test results (Table 3).

### 3.3 M-Layer Perceptron

The M-layer perceptron neural technique is efficient and fast computing methodology, uses feed forward network topology with minimal training datasets [5]. In general, M-layer perceptron consists of three layers as depicted in Fig. 1. The primary layer is input layer which has one or more neuron, the middle layer is hidden layer which has  $n + 1$  neurons, and the end layer is output layer which has one neuron. The mellitus datasets are taken as class instances and given as input dataset. The hidden layer does multiple analyze on the data and forms a pattern which is then grouped with similar datasets. This system of neural network algorithm is not defined with feedback signal for the error cancellation and noise reduction. The  $n + 1$  neurons in hidden layer determine the system stability and overfit under the RoC analysis. The accountability of neurons in the hidden layer cannot be performed. In this paper, concept of understanding and comparative analysis can be performed with six hidden neurons.

The datasets after processing with the hidden layer associated with the weight  $W_{(m,n)}$  along with the input  $a_m$  can be described as

**Fig. 1** M-Layer perceptron neuron architecture



Output = Input \* Weight associated with the hidden layer neurons

This can be represented as the function in equation

$$Z_m = F\left(\sum X_m W_{(m,n)}\right)$$

where  $f$  defines the activation function associated with the weighted inputs. It can be simple, tangential and trigonometric function to define the best results at the output layer. To initialize the weight of neuron in the hidden layer, the weights of the neuron will be assumed to be small in numbers in which it has been multiplied with the input neuron for the processing.

The hidden layer has multiple neurons in which one layer has the weight of any random numbers which defines the subset formation. The weight of hidden layer neurons need not be the same, and it can differ at any layer. The output layer is a combination of both the input layer and the hidden layer associated with the weights of the neuron.

The M-layer perceptron is a feed forward system in which the signal feedback to enhance the noise signals. The input of the second layer of neurons can be calculated using the formula

$$P_m = \sum_m W_{mn} Z_m + \theta_m$$

where the  $P_m$  defines the output from the previous layer which is taken as an input to the second level;  $W_{mn}$  defines the weight of the neurons associated in the hidden layer;  $Z_m$  defines the output neuron;  $\theta$  defines the bias function of the system. The M-layer perceptron results in the highest accuracy when compared to any other neural algorithmic technique due to multiprocessing on the input system.

### 3.4 Nearest Neighbor Neural Algorithm

Nearest neighbor neural technique is a classification methodology requires memory for computation and classification [6]. The algorithmic technique uses the memory to behold the data values which are computed and analyzed in the series of array in which all the datasets are stored and compared with each other to find the nearest neighboring point. Once the nearest neighbor is found by the nearest possible integer, the existing memory will be updated with the new memory, and again, the computational iteration will be carried out. The number of iterations in this mode of prediction methodology requires huge memory space to store the buffer values. The buffer values are nothing but the values that are not nearest to the main memory value. The number of iterations, memory space and buffer values are found to be huge for this computational mechanism.

#### Algorithm for Nearest Neighbor

**Step 1:** The dataset values are computed with the nearest neighbor neural technique by means of feed forward and classification of datasets based on the similarity in pattern formation and recognition.

**Step 2:** The output value will be stored in the main memory till the comparative analysis is been done with the next subset value.

**Step 3:** If the main memory value is lesser than the data subset value, then the main memory value will be replaced with the subset value which it means the first nearest neighboring node is found and updated.

**Step 4:** If the value is greater, then the main memory value retains the same, and the subset value is recomputed and compared. The process is continuous from Step 3.

**Step 4:** If the validation of Step 3 is correct, then the distance between the neighbor nodes can be found by using the equation mentioned above. The node distance can be found by comparing the main memory value and data subset value. Equation to calculate the distance of the neighboring points and the main memory behold value is specified below.

$$\sqrt{\sum_{i=1}^p (m_i - n_i)^2}$$

where  $m_i$  defines the main memory value,  $n_i$  defines the computed value. The iterations are performed by the limits from 1 to  $p$  which is probabilistic in nature.

### 3.5 Vector Machine

Vector machine is a classification methodology used in neural algorithmic technique for pattern formation and classification [7]. This system generalizes the datasets and forms patterns of similar and dissimilar class instances. This type of neural technique supports string and float datasets. Vector machine is to judge the datasets and define the specificity of datasets and their associated class instances. The ideology separates the patterns, and during this separation process, the vector machine collects the input data and matches the class instances with the data inputs to identify the similarity in patterns. The patterns are formed in cluster and plotted in space where the similar patterns are formed as group of cloud in the space and dissimilar patterns are observed separately from the cloud. The planes are constructed in multidimensional plot space which separates the class instances. Vector machine can handle regression and association with classification methodology which can hold the continuous and discrete datasets with category. The category parameters are created with the dummy parameters with either positive or negative instances.

$$A(1, 0, 0); \quad B(0, 1, 0) \quad \text{and} \quad C(0, 0, 1)$$

The hyper-planes are constructed with iteration which minimizes the error class instances. The error class instances are classified into groups by the vector machine [8]

1. Vector Machine by Classification (C-CVM)
2. Vector Machine by Classification-2 ( $\mu$ -CVM)
3. Vector Machine by Regression-1 ( $\epsilon$ -RVM)
4. Vector Machine by Regression-2 ( $\mu$ -RVM)
  
1. Vector Machine by Classification (C-CVM)

In this type of classification system, the error class instances can be minimized by the equation,

$$\frac{1}{2}C^T C + W \sum_{i=1}^p \varepsilon_i$$

where  $W$  is the constant of the class instance capacity and  $C$  is the coefficient of vector machine matrix. The  $\varepsilon_i$  determines the parametric constants associated with the inputs. The iterations are performed to increase the accuracy of the prediction methodology and pattern formation and recognition. It also enhances the conversion methodology to minimize the noise signals into useful similar pattern system.

2. Vector Machine by Classification-2 ( $\mu$ -CVM)

This method of classification system which is contrast to the classification methodology-1 is further enhanced to minimize the error.

$$\frac{1}{2}C^T C - l\rho + \frac{1}{p} \sum_{i=1}^p \varepsilon_i$$

The system defines with the increased iteration to minimize the noise datasets.

### 3. Vector Machine by Regression-1 ( $\varepsilon$ -RVM)

The noise signal is a function of input signals which are related and equated to the output datasets. This can be represented as

$$Z = f(m) + N$$

The function of input signals identify the new conditions on this prediction methodology. The vector machine achieved for training the system with the sample sets and the error functions.

### 4. Vector Machine by Regression-2 ( $\mu$ -RVM)

$$\frac{1}{2}C^T C - d \left( l\rho + \frac{1}{p} \sum_{i=1}^p \varepsilon_i \right)$$

The vector machine regression system of methodology enhances the further regression system by eliminating the noise signals and increases the accuracy of the prediction methodology.

## **3.6 Insulin Levels and Their Indications**

The insulin level analyzed from the algorithmic techniques indicates their status for the intensive care and generalized medication purpose. The forecast methodology finds to be accurate in indicating the status of the insulin level with the postprocessing state [9].

The various disorders of diabetic disorders are

1. Polyurea—results in excessive urination
2. Polydipsea—results in excessive thirst
3. Polyphagea—results in excessive hunger and increased secretion of hydrochloric acid.

## **4 Results and Discussion**

The datasets are taken from the blood test analysis of a person, and this value has been utilized for the neural algorithmic technique for processing the data [10]. The

**Table 4** Insulin level and their indication status

Parameters	Patient data	High risk	Intermediate risk	Risk range	Intermediate risk range	Optimal risk range
Glucose	70		100	>120	90–110	60–89
HbA1	4.2	8.2	—	>6	4.9–5.5	<4.5
Fructosamine	320	385	—	>340	290–332	<290
Glycation Gap	0.35		—	>0.7	0.30–0.45	<0.30
Post-Meal Insulin Level	5.5	15	—	>7.5	5.0–6.9	<5.0
Fatty Acid	0.49	0.40	—	>0.5	0.4–0.7	<0.4
Oleic Acid	60	80	—	>75	50–69	<50
Linoleoyl Acid	9.2	8.1	—	<10	9–10.5	<8.9
Insulin	9.2	10	—	>10	9–10	<9
Pro Insulin	7.2	7.9	—	>12	7–14	<6.9
Peptides	3.0	—	0.7	>4.2	2.9–3.7	<2.8

**Table 5** Accuracy of M-layer perceptron

	Instances	Percentage (%)
Similar instances	79	79
Dissimilar instances	21	21
Total	100	100

class instances are referred to as parameters, and the patient database is taken into consideration. But for the preliminary testing, a patient result has been taken into consideration along with the baseline values and the risk factor values (Table 4).

The patient data are taken into the neural algorithmic process state, and it is first processed with M-layer perceptron along with the datasets resulted by the National Family Health Survey Committee. The results are displayed in Table 5.

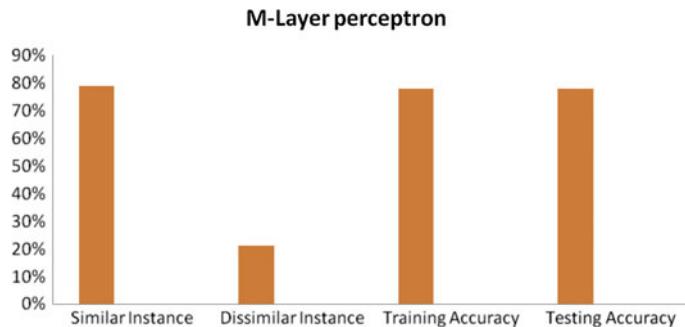
The training accuracy of the system is found to be 79.02%, and the testing accuracy is found to be 77.9% which is depicted in Fig. 2.

By means of nearest neighbor algorithm, the class instances are obtained from the patient data obtained in real time along with the mellitus datasets from National Health Survey committee also processed in the system to analyze the accuracy and specificity (Table 6).

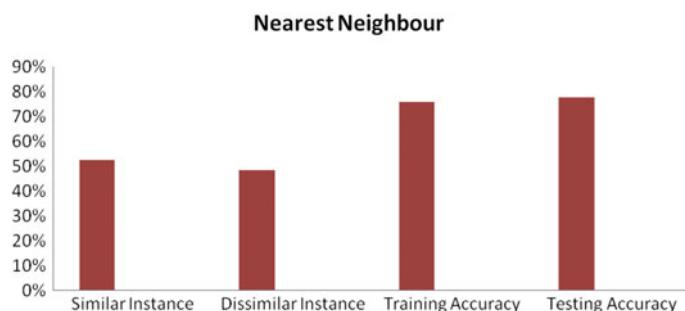
The training accuracy of the system is found to be 76%, and the testing accuracy is found to be 74%. The absolute error of the system is found to be 21% which is due to the dissimilar pattern formation and recognition (Fig. 3).

The vector machine system of classification is also analyzed with the datasets provided in real time and also with the dataset provided from National Family Health Survey Committee. The results are given in Table 7.

The training accuracy of the system is found to be 72%, and the testing accuracy is found to be 71.9%. The results are depicted in the graph shown in Fig. 4.

**Fig. 2** M-Layer perceptron training accuracy and testing accuracy**Table 6** Accuracy of nearest neighbor algorithm

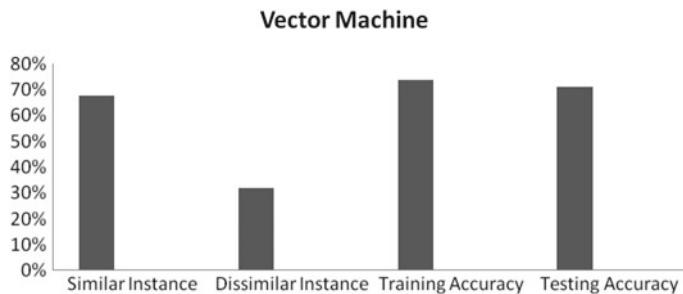
	Instances	Percentage (%)
Similar instances	51.69	51.69
Dissimilar instances	48.31	48.31
Total	100	100

**Fig. 3** Nearest neighbor training accuracy and testing accuracy**Table 7** Accuracy of vector machine

	Instances	Percentage (%)
Similar instances	68	68
Dissimilar instances	32	32
Total	100	100

The datasets are iterated to the value, and the accuracy, specificity and the sensitivity are analyzed with the respect to the iteration.

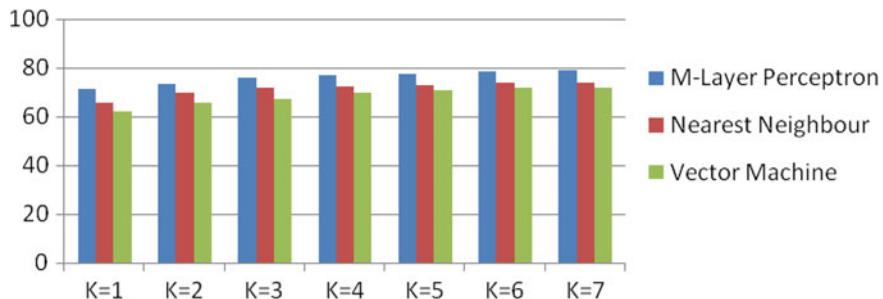
Table 8 defines the iteration of algorithm, and their accuracy values are determined at each stage of iteration.



**Fig. 4** Vector machine training accuracy and testing accuracy

**Table 8** Accuracy of classifiers

Algorithm	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
M-Layer perceptron	71.26	73.39	75.83	76.71	77.49	78.21	79
Nearest neighbor	65.54	69.98	71.63	72.29	72.63	73.79	74
Vector machine	62.30	65.78	67.41	69.92	70.74	71.82	72



**Fig. 5** Comparative analysis of accuracy along with iterative measurement

The results show for iteration  $k = 7$ , the training accuracy is found to be equal to the theoretical analysis, and the graph depicts the accuracy of system at various levels of  $k$  (Fig. 5).

## 5 Conclusion

In this classification methodology of supervised learning system, the M-layer perceptron, nearest neighbor and vector machine neural algorithmic technique are analyzed with the real dataset and the dataset obtained from National Family Health Survey

Committee. The mean factor and the distribution of datasets based on various classification conditions such as person behavior, mellitus disorder type and the forecast result over a period of time performed in this system. The M-layer perceptron results found to have accuracy compared with any other methodology due to feedback system which rectifies the deviation from the original dataset to the erroneous dataset. The accuracy of M-layer perceptron is found to be 79%, and it is same in the training phase, and in testing phase, it is found to be 77%. By change in the value of key attributes, the intensive caring period is avoided by the forecast methodology by improvising the noise cancellation and elimination of the processed datasets. The noise elimination is done by means of feedback signals which reduce the error by 0.8%.

## References

1. <https://www.webmd.boots.com/diabetes/types-diabetes-mellitus>
2. Selvakumar S, Kannan S (2017) Prediction of diabetes diagnosis using classification based data mining techniques. *Int J Stat Syst*
3. Soltani Z, Jafarian A (2016) A new artificial neural networks approach for diagnosing diabetes disease type II. *Int J Adv Comput Sci Appl* 7
4. [https://www.diabetes.com/diabetes\\_care/blood-sugar-level-ranges.html](https://www.diabetes.com/diabetes_care/blood-sugar-level-ranges.html)
5. Pradhan M, Kumar R (2011) Predict the onset of diabetes disease using artificial neural network. *Int J Comput Sci Technol* 2
6. Rahimloo P, Jafarian A (2016) Prediction of diabetes by using artificial neural network logistic regression statistical model and combination of them. *Bull Soc Sci Liege* 85
7. Sapon MA, Ismail K, Zainudin S (2011) Prediction of diabetes by using artificial neural network. In: International conference on circuits, systems and simulation, vol 7
8. Durairaj M, Kalaiselvi G (2015) Prediction of diabetes using soft computing techniques—a survey. *Int J Sci Technol Res* 4
9. M Kumari, R Vohra, A Arora (2014) Prediction of diabetes using Bayesian network. *Int J Comput Sci Inf Technol* 5
10. Shanker M (1996) Using neural networks to predict the onset of diabetes mellitus. *Int J Chem Inf Comput Sci*

# Nipah Virus Using Restricted Boltzmann Machine



Velpula Sandhya Rani, Havalath Balaji, Vishal Goar,  
and N. Ch. Sriman Narayana Iyengar

**Abstract** Nipah virus is an infectious virus which is caused by fruit bats. Recently in 2018, there was a deadly outbreak that occurred in Kerala where many of people got infected and died due to Nipah virus. In this model, we are using deep learning concept which helps to predict the occurrence of infected virus using restricted Boltzmann machine. It is a feature selection algorithm where particular data will be selected by applying matrix of weights associated with the connection between the hidden layer and the visible layer. Firstly, it was identified in Malaysia Kampung Sungai Nipah in 1998. The fertility rate people affected with Nipah virus was around 70%. Transmission of this infected virus is done by bats-to-human, animals-to-human and human-to-human. Particular signs and symptoms will be exhibited for the person affected with Nipah virus. Cerebrospinal fluid serum test will be done by collecting white blood cells, glucose and protein by using restricted Boltzmann machine. This deep learning algorithm will give the numeric results such that we can identify whether the patient is affected with Nipah virus. Prevention measures should be taken as “prevention is better than cure” because there is no vaccine for Nipah virus which is eventually more dangerous.

**Keywords** Deep learning · Nipah virus · Restricted Boltzmann machine

---

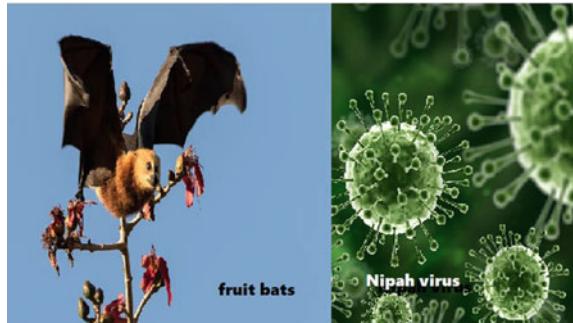
V. S. Rani · H. Balaji · V. Goar · N. Ch. Sriman Narayana Iyengar (✉)  
Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India  
e-mail: [srimannarayananach@sreenidhi.edu.in](mailto:srimannarayananach@sreenidhi.edu.in)

V. S. Rani  
e-mail: [sandhyavelpula4@gmail.com](mailto:sandhyavelpula4@gmail.com)

H. Balaji  
e-mail: [balajimitk@gmail.com](mailto:balajimitk@gmail.com)

V. Goar  
e-mail: [dr.vishalgoar@gmail.com](mailto:dr.vishalgoar@gmail.com)

Government Engineering College, Bikaner 334001, India

**Fig. 1** Nipah virus fruit bat

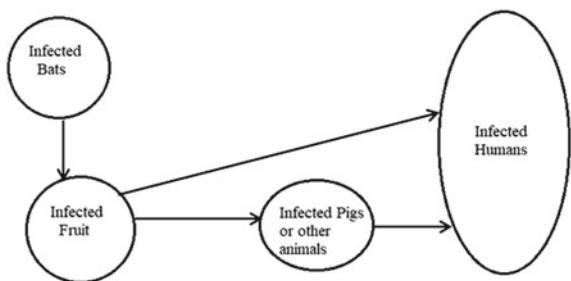
## 1 Introduction

Nipah virus is a zoonotic virus which is caused due to infected bats also called as flying fox or fruit bats. Zoonotic means the transmission of diseases from vertebrate's animals to humans. Nipah virus belongs to the genus of Henipaviral [1]. Nipah virus was firstly identified in the year 1999 in a village known as Kampung Sungai Nipah during the outbreak among the pig farmers in Malaysia, and later, it was also recognized in Bangladesh in 2001. Recently, it was also identified in Kerala unexpectedly where there was a major outbreak, and people were infected by consuming date palm sap that was contaminated by virus-carrying bats. Nipah has been occurred last 15 years ago but identified in 2001 Siliguri, 2007 Nadia, and recently in 2018, there was a major outbreak in Kozhikode and Malappuram district of Kerala where many people were affected with Nipah virus. Flying bats can be found in Asia, East Africa and Pacific Islands [2]. The major outbreak occurred in Malaysia, Bangladesh and India. Fruit bats are the natural hosts which transmit deadly diseases along with Ebola and Zika. Nipah virus is carried by fruit bats, and it does not make bats sick because it has high metabolic rate and high body temperature. There immune system will fight against possible infections that occur. Some bats were infected from bats saliva where the transmission of virus is done from one to another which is very harmful spreadable disease. Virus may infect to human if they consume a fruit nibbled with infected bats [3]. In this paper, we are analysing the data of patients who are infected with Nipah virus (Figs. 1 and 2). The data consists of cerebrospinal fluid test details like white blood cells count, red blood cells count, glucose level and protein level.

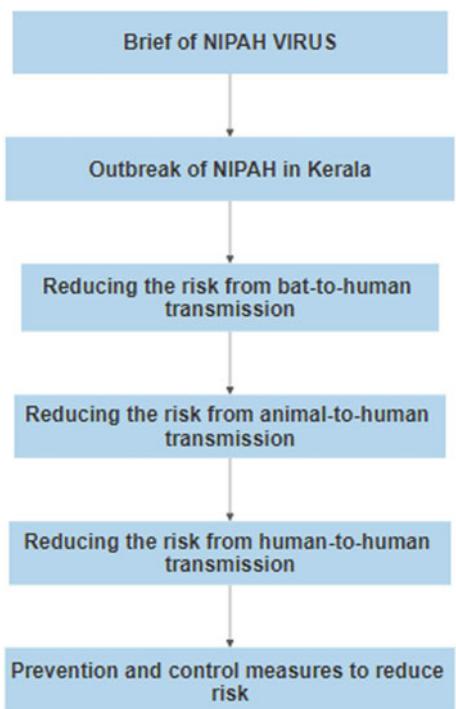
## 2 Transmission of Nipah Virus

Firstly, it was recognized in the outbreak of Malaysia, and later in Singapore, humans were affected by contact or body fluids such as blood, urine or excreta of flying fox on fruits. The process of reducing transmission of Nipah virus is given in Fig. 3.

**Fig. 2** Spreading of Nipah virus



**Fig. 3** Reducing transmission risk of Nipah virus



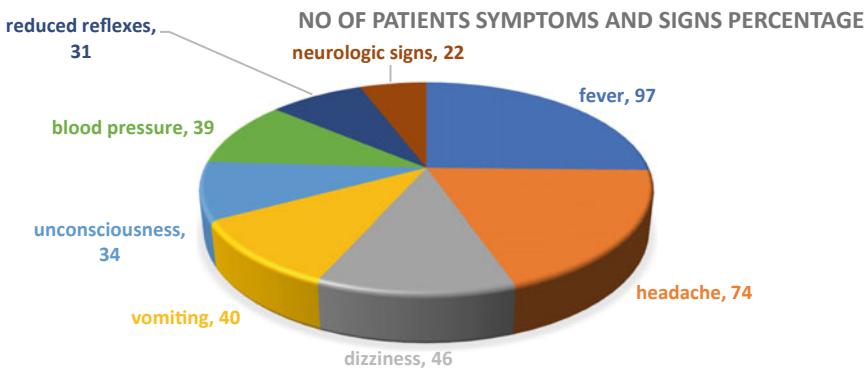
Pigs or other animals consume these infected fruits where the occurrence of Nipah virus takes place or also by unprotected contact with the tissue of infected animals [4]. There is no proper vaccine, so we can reduce the risk factor by educating the people about the control measures and prevention. We can diminish the risk of spreading virus from bats-to-human by only collecting the fresh raw date palm juice or washed fruits before consumption. Fruits with sign of bats should be avoided where circumstances of transmission virus will be more. We can reduce the risk from transmission from animals-to-humans by wearing protective gloves or clothes while handling sick pigs [5]. We should avoid contact with sick animals as much as possible. We can also reduce the risk from human-to-human transmission by avoiding physical contact with

infected people, and we should also ignore using their towels or their belongings. We should regularly wash our hands after taking care or visiting sick people.

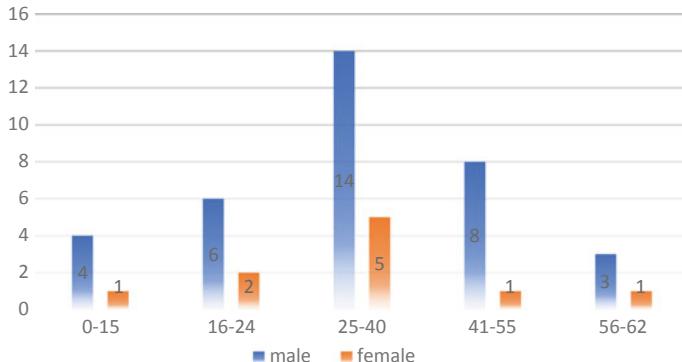
World Health Organization (WHO) issued guidance to take all the prevention methods by State Government of Kerala along with Central Minister of Health and Family [4]. They are supporting for the people getting affected or the risk of transmitting virus by technical guidance. Infected people initially have symptoms or particular signs such as fever, headache, vomiting, dizziness, unconscious, blood pressure, reduced reflexes and neurological signs. Incubation period is of 4–14 days after which the person can exhibit symptoms. This virus is similar to Influenza or Flue. Across 20% of people infected with Nipah virus and having acute encephalitis makes recovery, but long-term neurological conditions have been occurred by analysing the previous reports. Asymptomatic infection may also occur in some cases where the person will just carry the virus without having any symptoms [6]. People affected with Nipah virus have high fertility rate of about 70% to die. From a long period of time, the consequences of virus have been noted that personality changes also transpire. Death has been raised after some months and even years after the signs and symptoms have been took place for the people infected with Nipah virus.

The percentage of occurrence of fever will be of 97% which will be common for most of the people affected with Nipah virus, neurological signs are about 22%, reduced reflexes are about 31%, blood pressure is about 39%, becoming unconscious is about 34%, vomiting 40%, dizziness 46%, and headache is about 74% for most of the people. If the person is affected with any of the symptoms, then we can justify that person is infected. These are the major symptoms that occur for the people who got infected with Nipah virus. Symptoms along with number of patients is given in Fig. 4.

If we observe Fig. 5, the people affected with Nipah virus are more in male rather than female. From age 25–45, the occurrence of Nipah virus is more than other age people.



**Fig. 4** Symptoms and signs



**Fig. 5** Male and female patients affected with Nipah virus

### 3 Deep Learning

**Deep learning** is a part of machine learning. The data will be transformed into number of layers, where each layer will perform the specific process [7]. Deep learning is usually achieved by utilizing neural system engineering. In deep learning, training and testing of the data will be done to predict the best result. We can focus on the prediction of health issues by using the learning algorithms. Learning process can be done in both supervised and unsupervised manner. Big data and high-performance maintenance can be done by using deep learning.

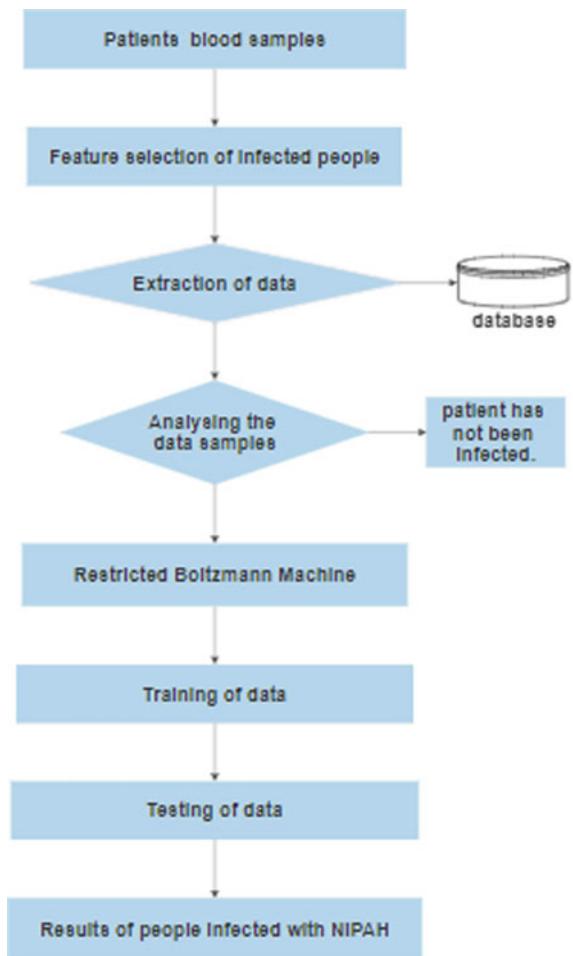
### 4 Proposed Work

We use deep learning for implementation through high-level programming. Transmission of data is done between numbers of layers such that it can predict to generate accurate output. Here, the performance is done from one layer to another in a specific manner such that redundancy can be removed in the layered structure. Many works have been done to detect high yield techniques for remedy of finding different diseases. Similarly, here we are checking whether the particular patient is affected with Nipah virus or not by using restricted Boltzmann machine (RBM). It has a structure like bipartite graph where intra communication between two nodes is maintained in a layer.

## 5 Methodology

Firstly, the feature selection is done based on the weights provided to different symptom in Nipah virus by using neural networks. Those features are extracted by using the data values applied to restricted Boltzmann machine for classification [8]. The process model consists of data where feature selection and extraction will be done by using restricted Boltzmann machine. Appropriate linkage will be formed between each layer and predict the result by historical information. The figure mentioned below detects whether the candidate is affected with Nipah virus or not by extracting the data in a particular manner (Fig. 6).

**Fig. 6** Dataflow diagram by using restricted Boltzmann machine



**Data Preprocessing:** In data preprocessing, steps will be followed where the selection of only three features will be done from a file [9]. Then, the normalization of data will be done by following min max normalization such that we can get the input vector range between 0 and 1 to avoid complexity. The neighbouring dataset will diverge into training dataset and test dataset such as 20% will be of test data and the remaining 80% of training dataset. The upper bound (UB) preference is 1, and the lower bound preference is 1. To find min max normalization of data, we use Eq. (1)

$$X_{\text{norm}} = (X - X_{\min}) / (X_{\max} - X_{\min})(\text{UB} - \text{LB}) \quad (1)$$

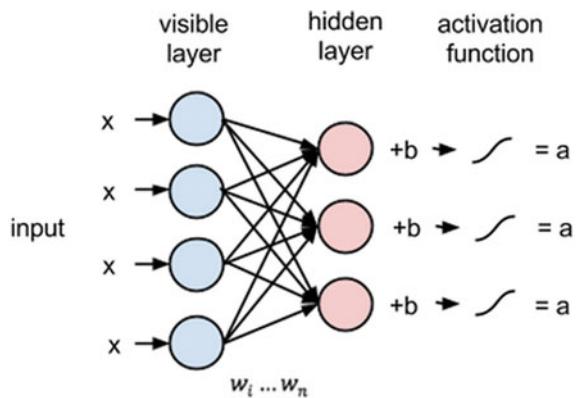
**Restricted Boltzmann Machine** is model which is having structure like bipartite graph, and it is energy-based model. Its graphical structure has undirected graph. It has visible layer and hidden layers by adding weights. The input values are directly assigned with the visible layer. The real valued data can be applied to RBM which is having a similar structure as Gaussian Bernoulli type architecture [10]. Modified RBM with binary logistic hidden units and the real values of Gaussian visible units can be used in classification (Fig. 7).

A restricted Boltzmann machine is a definite kind of a Markov random field which consists of two layers [11]. One layer of an RBM comprises visible input units,  $v$ , which are associated with the other layer of shrouded stochastic units is  $h$  either 0 or 1. The basic structures demonstrate the essential access of data. The appropriation of state  $\{v, h\}$  of a RBM is indicated by the accompanying vitality work.

$$E(v, h) = \sum_j \frac{(vi - bi)^2}{\sigma_i^2} - \sum_j \frac{vi}{\sigma_i} * w_{i,j} * h_j - \sum_j c_j h_j \quad (2)$$

$W$  is the weights applied represents from visible to hidden layer comprising as  $w_{ij}$  of associations between neurons  $v_i$  and  $h_j$ , and  $b$  represents to a visible bias vector. The positioning of all the parameters can be indicated by  $\theta = \{W, b, a\}$ .

**Fig. 7** Basic structure of RBM



Joint probability distribution is computed using formula (2) to obtain maximum probability. The distribution computation energy is kept minimum as the negative energy increases in probability and vice versa

$$P(v, h) = \exp(-E(v, h))/Z \quad (3)$$

where  $Z$  is partition function by summing all the pairs of visible and hidden unit given by below equation.

$$Z = \sum_{v,h} e^{-E(v,h)} \quad (4)$$

The conditional probability distribution of each unit is given by the sigmoid activation function of the input it receives using below formula:

$$P(h_j|v) = \text{sigm}\left(\sum_i w_{i,j} v_i + c_j\right) \quad (5)$$

$$P(v_i|h) = \text{N}\left(\sum_j w_{i,j} h_j + b_i, \sigma_i\right) \quad (6)$$

$$\text{Sigm}(x) = 1/(1 + \exp(-x)) \quad (7)$$

Computing  $p(r, x)$  is unmanageable, but it is possible to compute  $p(r|x)$  samples from it or we can choose the most likely happening class under this model. Reasonable number of classes can be computed exactly and efficiently by conditional distribution  $C$ . To compute energy distribution when there is classification-related problem, we use the below formula, distribution of energy function as:

$$P(v, h, r) = \exp(-E(v, h, r))/Z \quad (8)$$

Using the distribution given in Eq. (8), the classes are predicted such as to check patient has been infected or not.

Steps in training restricted Boltzmann machine are

1. Take the training data set directly to visible unit.
2. To update the hidden states, use sigmod activation function equation number 7.
3. For  $i$ th hidden unit, compute activation function using equation number 5.
4. Set the visible unit value to 12 using formula 3 and unit value to 0 using Eq. 6.
5. Compute the positive statistics for edge  $(e_{ij}) = v_i * h_j$ .
6. Again, reconstruct the visible unit using the similar technique. For each visible unit, compute the activation energy using Eq. 5 and update the state.
7. Now, update hidden units again, and compute  $(e_{ij}) = v_i * h_j$  which is the negative statistics for each stage.

**Table 1** Cerebrospinal fluid range of a normal person

1.	WBC count, cells/mcL	4000–11,000
2.	Protein level, mg/dL	20–40
3.	Glucose level, mmol/L	2.5–4.4

In restricted Boltzmann machine, the visible units are used to find the distribution of hidden units, and the hidden units are used to compute the distribution of visible units until the required stable state has been obtained. If the people have the above symptoms and signs, the person should consult doctor and check cerebrospinal fluid serum test where the blood samples were collected and check the count of white blood cells, glucose level and protein level.

If the count is more, then we can justify that the person is infected with Nipah virus. Cerebrospinal fluid test has been examined to detect any disruption that occurs in our brain. The cerebrospinal fluid test consists of the level of our glucose, protein and white blood cells count. If the people have the above symptoms, check cerebrospinal fluid serum test where the blood samples were collected and check the count of white blood cells, glucose level and protein level [12]. If the count is more, then we can justify that the person is infected with Nipah virus. The cerebrospinal fluid is filled in ventricles in human body. Brain controls our entire body system where it is surrounded with CSF fluid (Table 1).

## 6 Performance and Evaluation

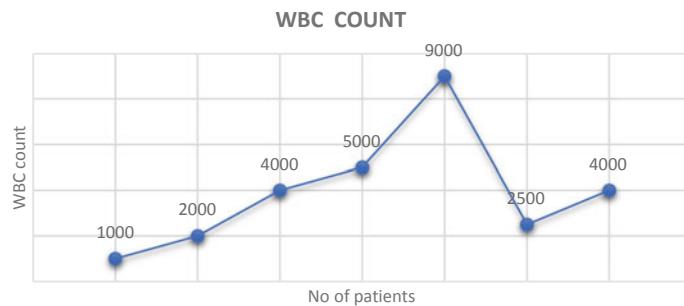
**White blood cells** (WBCs) are the cells of immune system. They help to protect our body against infectious diseases as well as foreign invaders. WBC is also called as leukocytes. The bone marrow continuously produces WBC until they fight against any infectious diseases. The normal range of WBC count must vary between 4000 and 11,000 cells/mcL.

The persons affected with Nipah virus have very less WBC count (Table 2; Fig. 8).

If you observe the above graph having WBC count is less in some patients, we can identify the sign that the patients having more WBC have been infected. This is due to WBC generated by bone marrow which is a spongy tissue inside some of our larger bones.

**Table 2** WBC count of people infected with Nipah

No of patients	1	2	3	4	5	6
WBC count cells/mm <sup>3</sup>	1000	2000	4000	5000	9000	2500



**Fig. 8** Shows the content of WBC count of people infected with Nipah

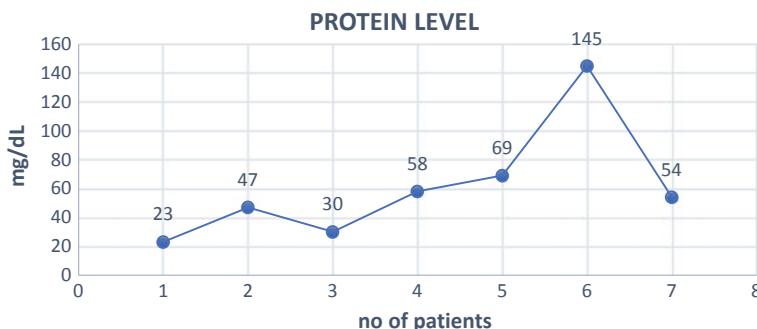
**Proteins** are the requisite nutrients for human body. They are one of the building blocks of our cell tissue which is like a fuel source. Energy density of proteins is more. Proteins are the large complex molecules which consist of amino acids (Table 3; Fig. 9).

Doctor cleans our back with antiseptic and applies a local anaesthetic. This numbs the puncture site to minimize pain. It may take a few moments to start working. Then, they insert a needle into our lower spine, and they extract some small amount of CSF into the needle. Then, the doctor removes the needle after accumulating adequate fluid from the body [13]. They clean and dress up the site where they have inserted the needle. Later, they send CSF sample to laboratory for analysis.

**Glucose** test is done in cerebrospinal fluid test, and we will measure the aggregate glucose level or sugar level present in the fluid of a human body. People infected with

**Table 3** Protein level of people infected Nipah

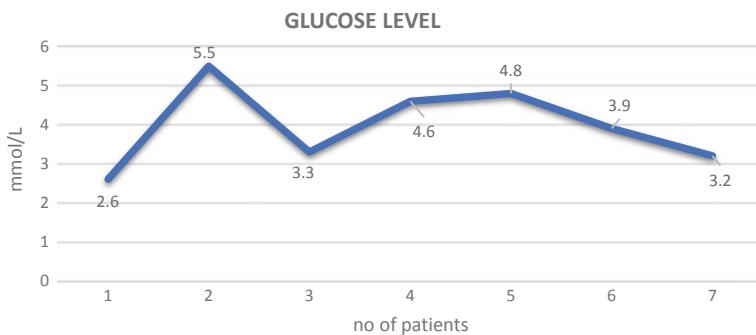
No of patients	1	2	3	4	5	6	7
Protein level mg/dL	23	47	30	58	69	145	54



**Fig. 9** Shows the content of protein level of people infected Nipah (Table 3)

**Table 4** Glucose level of people infected with Nipah

No of patients	1	2	3	4	5	6	7
Glucose level mmol/L	2.6	5.5	3.3	4.6	4.8	3.9	3.2

**Fig. 10** Shows the glucose level of people infected with Nipah

virus will have high glucose level as their blood gets infected by the consumption of infected food which is converted into glucose. The level of glucose may be high because it does not often show any signs. Infected blood carries throughout every cell present in our body.

CSF is a clear fluid surrounded with brain and spinal cord is the better course of action for testing the conditions because it is in direct connection with brain and spine. We will examine all the above results and monitor whether the particular person is affected with Nipah virus or not in a particular order by using our algorithm (Table 4; Fig. 10).

## 7 Discussion

WBC count ranges from 1000 to 9000 normally, the WBC count must vary from 4000 to 11,000, but for some patients, the WBC count is less such that it can be identified that the particular patient has been affected with some infection. Normally, the level of protein should range from 20 to 40 mg/dL. The above graph shows the protein level of some patients is more which are the signs of tumour or some occurrence infections. The glucose level test helps in identifying the conditions that had been raised in our body such as infections and tumours. All the above values will be analysed and the restricted Boltzmann machine algorithm will check whether the patient has been infected with Nipah virus or not.

## 8 Prevention

First acupuncture for human beings is intensive support care (ISC) by WHO for the people affected with Nipah virus [10]. The ISC focuses on symptoms such that they can remove the infection by taking care of the symptoms that affected people with Nipah virus. Personal protection equipment (PPE) should be taken for the workers and the people who collect samples from patients. Contact tracing should be done because we can reduce the risk of virus transmission. Nipah virus can be averted by avoiding exposure of sick pigs or bats and also with direct contact with sick or unhealthy people. We should also avoid drinking raw date sap because “**prevention is better than cure**”. A centime vaccine, using the Hendra protein which provides defensive antibodies which has been lately used in Australia to protect horses from Hendra virus, but the result was not up to the mark. So, the government is trying to access the accurate vaccine to avoid Nipah virus.

## 9 Conclusion

Restricted Boltzmann machine in deep learning was effectively prepared for prediction of data. Nipah virus is caused by infected flying bats which had high metabolic rate. Recently, major outbreak occurred in Kerala where many people got infected. We are collecting the samples such that we can analyse and predict the occurrence of our infection that occurred. People living in rural areas do not understand about the infection, but by symptoms, they can analyse and consult doctor for sample test which can be done rural as well as urban areas. Transmission of Nipah virus for human is mostly by date palm sap which had been consumed by infected bats. Understanding of these diseases could lead to the breakage of treatment where prevention is done. Here, we should avoid contact with infected pigs and animals which are sick, and certain prevention measures should be taken. We will get the satisfactory outcome where the prediction of infectious can be identified using the measure which National Centre for disease control is presently using in Kerala for reviewing the situation occurred by bats.

## References

1. [outbreaknewstoday.com/nipah-virus-introduction-28899/](http://outbreaknewstoday.com/nipah-virus-introduction-28899/)
2. <https://www.who.int/csr/don/07-august-2018-nipah-virus-india/en/>; [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(18\)31252-2](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31252-2)
3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784122/> Transmission of human Infection with Nipah Virus
4. Chakraborty A, Sazzad HM, Hossain MJ, Islam MS, Parveen S, Husain M et al (2016) Evolving epidemiology of Nipahvirus infection in Bangladesh: evidence from outbreaks during 2010–2011. *Epidemiol Infect* 144:371–380

5. <https://www.sciencedirect.com/science/article/pii/S0020025518304870>
6. World Health Organization (2005) WHO publishes list of top emerging diseases likely to cause major epidemics
7. Chong HT, Tan CT (2003) Relapsed and late-onset Nipah encephalitis, a report of three cases. *Neurol J Southeast Asia* 8:109–112
8. proceedings.mlr.press/v5/salakhutdinov09a/salakhutdinov09a.pdf
9. Treatment | Nipah Virus (NiV) | CDC. Available from: <https://www.cdc.gov/vhf/nipah/treatment/index.html>
10. deeplearning.net/tutorial/rbm.html Restricted Boltzmann machine
11. outbreaknewstoday.com/nipah-virus-introduction-28899/
12. <http://www.who.int/medicines/ebola-treatment/WHO-list-of-top-emerging-diseases/en/>. Accessed 17 Aug 2018
13. MohdNor MN, Gan CH, Ong BL (2000) Nipah Virusinfection of pigs in Peninsular Malaysia. *Rev scitech Off intEpiz* 19:160–165

# Big Data Analytics—Analysis and Comparison of Various Tools



Amit Gupta, Bhanu Prakash Dubey, Himani Sivaraman, and M. C. Lohani

**Abstract** Big data is the latest terminology in the computer world. The data collection is increasing day by day, and many technological changes can bring some new methods for decision-making process in many areas such as health and finance. As the complexities are increasing due to volume, veracity, variety and velocity, our focus is on the methods to calculate the value of data using various big data analytics technologies. The analytics process used with respect to big data can be improvised by using new algorithms, which enhance the analytical aspects and can be used to extract the hidden knowledge very efficiently and effectively.

**Keywords** Big data · Hadoop · HDFS · Spark · Map reduce

## 1 Introduction

Big data terminology is used for the collection of various data sets which are diverse in format and complexity. Due to its diversity, these huge data sets are very difficult to be stored and processed using traditional data processing tools or applications. Thus, we require some techniques or concepts with the help of which we can easily work on and use these data sets for various purposes. Big data analytics facilitate the collection of data from different sources, transforming them to such a format so that it becomes ready to be used by various analysts and eventually providing it to various organizations. Big data and machine learning altogether enhance the performance of

---

A. Gupta (✉) · B. P. Dubey · H. Sivaraman · M. C. Lohani  
Department of CSE, Graphic Era Hill University, Dehradun, India  
e-mail: [amitgupta7920@gmail.com](mailto:amitgupta7920@gmail.com)

B. P. Dubey  
e-mail: [bhanu.dubey@gmail.com](mailto:bhanu.dubey@gmail.com)

H. Sivaraman  
e-mail: [himanisivaraman@gmail.com](mailto:himanisivaraman@gmail.com)

M. C. Lohani  
e-mail: [getmlohani@gmail.com](mailto:getmlohani@gmail.com)

various industries like finance, healthcare, etc. This is because the price of data storage has been reduced and accessibility to high end and high performance computer becomes easy. Thus, various theoretical concepts of big data when implemented using machine learning tools give enhancements to many industries and business organizations.

Nowadays, the generation rate of the data is very fast. Approximately, around 90% of the data which is present in present world has been created in previous two years. In recent decades, the huge amount of data is been generated from various sources like:

1. Walmart handles more than 1 million customer transaction every hour
2. Popular social media platform Facebook uses, stores and analyzes around 30 plus petabytes of data which is all generated by its millions of users
3. Approximately, 48 h of new video are been uploaded to YouTube every hour
4. Amazon handles near about fifteen million user activity click per day that plays an important role for recommending various products to its customers
5. Various mail servers analyze around 294 billion emails to find the spam mails
6. Modern vehicles have more than 100 different types of sensors to monitor various things like fuel consumption, tire pressure, etc., and thus, every vehicle generates lots of sensor data that can be stored and processed on Cloud.

## 2 Big Data Characteristics

### 2.1 *Volume*

Volume means that the enormous information and data which is generated on daily basis increases in exponential, and this huge amount of data mainly represented in terabytes, petabytes or in some cases even in zetabytes. This information or data is so big that it cannot be handled, managed or controlled by using ancient methods or traditional methods of data managing techniques. For example, the size of data being generated by the interaction between humans and machines through various social media platforms.

### 2.2 *Velocity*

Velocity means the speed with which various sources will generate data on daily basis. This huge data is very enormous and continuous in nature. For example, on Facebook, there are around 1.03 billion active users daily which approximately increases around 22% each year. This concludes that how fast the number of users is increasing on social media platforms. These users are responsible for the fast growing of data on

**Fig. 1** Data with missing values

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

daily basis. Simply, if you can handle the velocity, you will be able to generate various insights and will be able to take decision based on this real time generated data

### 2.3 Variety

In big data, many different types of data sources basically responsible for different types of data eventually contribute for the formation of big data. The data generated from various data sources can be structured, semi-structured or unstructured in nature. Traditionally, data was mainly stored in excel and databases, but nowadays, the data is collected in various formats like images, audio, video, sensor data, etc., and hence, this variety of semi-structured and unstructured data mainly creates problems related to storage, collecting, extracting information and analysis of data.

### 2.4 Veracity

Veracity [1] means that the data is in doubt or is uncertain of data availability due to the incomplete data or inconsistent data (Fig. 1).

Many a time, the data can be messy and untreatable. As in big data, the data occurs in many forms, and therefore, the quality and accuracy always remain a big problem. The volume is always responsible for the lack of quality and accuracy of data.

### 2.5 Value

With the volume, velocity, variety and veracity, we need to discuss one more V related to big data, this is value. It is basically the usefulness of the data. The features

and functions of big data include security, storing, analysis, exploring, visualization [2], modification and transactions. In today's world, there are various technologies and techniques [3] which can be used along with big data to perform faster and efficiently. Parallelism increases the speed at which big data can be processed, and it also increases the analyzing capabilities of the data. The usage of distributing computing [4] systems can be used for the efficient processing of big data mainly in real-time manner.

The various technologies used in Big Data which are treated as best four Apache Big Data Frameworks are described briefly in the following section.

### 3 Apache Hadoop

Apache Hadoop [5, 6] is basically an open source framework, which is written in Java. It is fault tolerant and scalable framework which provides batch processing techniques to be used in efficient way. It performs better than any other technique as it is capable of processing large volume of different forms of data on a group of various commodity hardware. Hadoop is mainly misunderstood as a system to store data, but instead, it is a technology or method which possesses capability for storing large volume of data along with processing of large amount of data.

Hadoop is technology that is designed to process big data which is combination of both structured and semi-structured data which is available in huge volume. It also provides analytical techniques and computational power required to work with large and diverse form of data.

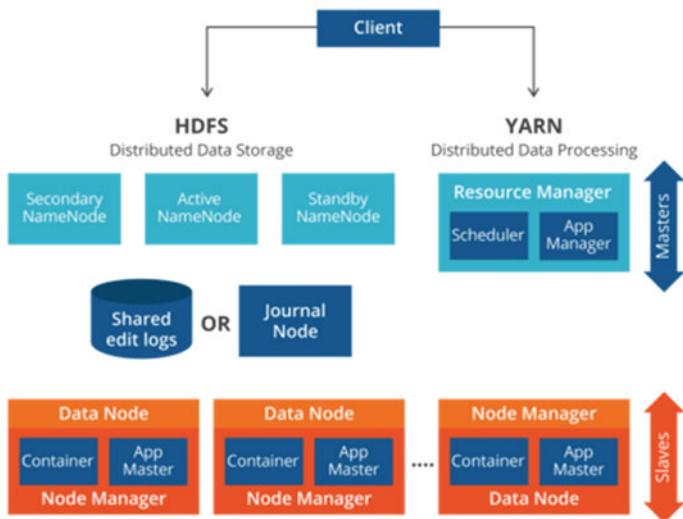
Hadoop framework is an example of cluster which comprises one master node and many worker nodes. This master node is composed of both Name Node and Job Tracker Node, whereas Worker Node can act as both Data Node—responsible for storage of data and Job Tracker—responsible for monitoring jobs. It also contains Secondary Name Node which is the replication of Name Node. The responsibility of the Secondary Name Node is to take snapshot of Primary Name Node directory information at regular interval of time. This can be used in place of Name Node to restart the faulty or failed Name Node (Fig. 2).

Hadoop consists of two main components: Hadoop Distributed File System (HDFS) and Map Reduce [7].

### 4 Hadoop Distributed File System (HDFS)

HDFS is used for storage and is fault-tolerant mechanism that stores large size files from terabytes to petabytes across different terminals in distributed manner. The default value of replication is 3 that can be increased according to the sensitivity of data being stored. It splits big file into large block size of 64 MB (can be changed to 128 MB) and can be stored independently on multiple nodes. Its main responsibility

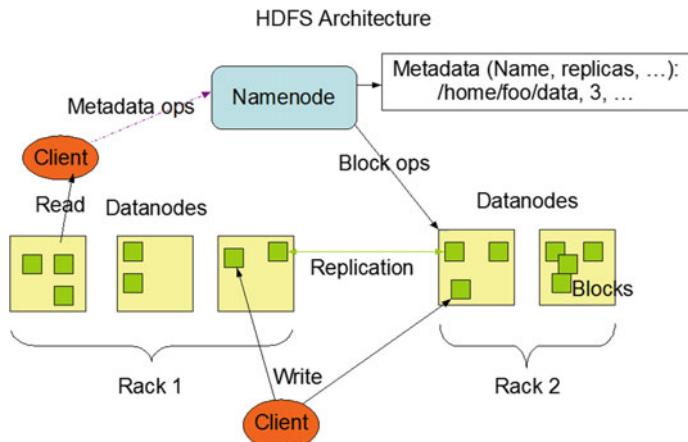
## Apache Hadoop 2.0 and YARN



**Fig. 2** Apache Hadoop and Yarn

is to ensure the availability of data even during the failures of the host machine. It is also used to store immediate processing results. HDFS is mainly suitable for the distributed storage and processing. Hadoop provides a command interface to interact with HDFS for the streaming access to file system data (Fig. 3).

HDFS provides an automatic fault detection mechanism which improvises its mechanism of recovery process during disaster. HDFS includes large number of



**Fig. 3** HDFS architecture. *Source* [hadoop.apache.org](http://hadoop.apache.org)

various hardware, and thus, failure of any component is an issue. Therefore, it provides an efficient recovery system to facilitate efficient working of the Hadoop system. The processing methodology of HDFS is such that it always selects node for processing local node to reduce network traffic and increase throughput.

## 5 Map Reduce

Hadoop Map Reduce [8, 9] is basically a software framework for easily providing various processing tasks that involves huge amount of data. It basically facilitates parallel execution of application on large clusters in fault-tolerant manner. Map Reduce [10, 11] is programming structural model for writing tasks which can be executed in parallel fashion on multiple nodes. It also provides analytical capabilities for complex data. Traditional model has a limitation as it cannot provide mechanism to process huge volumes of scalable data, and on the other hand, the centralized system provides too much of a bottleneck while processing multiple files simultaneously. Google has developed an algorithm to solve this issues, and this technique is called Map Reduce [11, 12] which divides the task into small part and assigns them to different nodes. After processing, these individual results are combined to give the integrated output. The Map Reduce algorithm consists of two important activities: Map and Reduce [5, 13]. Map converts the data sets into individual elements of key, value or tuple. Reduce collects the output from each mapper and combines them. The most important benefit of using Map Reduce is that it provides an easy mechanism and method to distribute data processing on multiple and different computing nodes.

## 6 Apache Storm

It is a framework which mainly focuses on low latency. It provides an efficient and better option for processing which actually requires real-time processing. It provides an efficient methodology that works on huge amount of data and reduces latency in comparison with other frameworks. Storm has facilities such as real-time analytics, online machine learning, continuous computation and ETL, and it is scalable, fault tolerant, guarantees efficiently processing of data. There are certain features that make storm more powerful tool rather than Hadoop like fault tolerant, scalable, fail fast, auto restart approach, support multiple languages and Json, support for direct acyclic graph (DAG) topology, etc.

## 7 Apache Samza

It is a stream processing framework that is strongly associated with Apache Kafka messaging system. It is designed specifically to enhance the benefits of Kafka's specific architecture. Like other technologies, it also uses fault-tolerant mechanism for buffering and storage. For the purpose of resource negotiation, it uses YARN [14] along with its rich features.

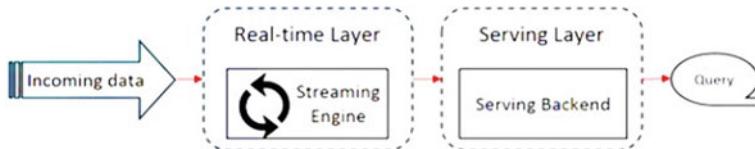
## 8 Apache Spark

Apache Spark [15, 16] is an example of all purpose cluster computing system which possesses huge and large number of libraries and APIs for various programming languages such as R, Python, Scala and Java. Unlike Hadoop, it is very fast and efficient in processing and accessing data from the storage. It can be implemented by using Hadoop or without Hadoop. It mainly focuses on quick execution of the task by implementing the methodology of batch processing workload in memory computation. It can be implemented as standalone cluster and can be used with Hadoop as an alternative to Map Reduce. The main component of Spark [17, 18] is driver program, cluster manager and worker node. The driver program is on the spark which starts the execution of any application. The cluster manager allocates all the resources. Lastly, Worker Node does all the processing. Some properties of the spark which makes it better than Hadoop are its high speed, high performance, high query optimization. It can run on any platform, has a large library set and data pipelining facility.

## 9 Apache Flink

Apache Flink is a platform which is categorized as open source; unlike any other framework, it has a flow engine for streaming data which also provides a methodology for communication, fault tolerant and distribution of data on various distributed computations over streaming data. This framework of data analytics is wholly compatible with Hadoop. Flink has the capability to execute both streaming processing and batch processing [19, 20] without any difficulty.

Because of the micro-batch architecture of spark, it is not suitable for many use cases. It is also enriched by the batch and stream processing capabilities. Apache Flink provides low latency, high throughput and real transactional processing. The architecture of Kappa forms the basis for working of Flink. The benefit of using Kappa architecture is it has only single processor—stream, which considers various input as stream, and the streaming engine present in the Kappa processes the entered



**Fig. 4** Apache Flink architecture

**Table 1** Comparison of various tools

	Execution model	Supported language	In-memory processing	Low latency	Fault tolerance	Enterprise support
Map reduce	Batch	Java	No	No	Yes	No
Storm	streaming	Any	Yes	Yes	Yes	No
Spark	Batch and streaming	Java, Scala, Python, R	Yes	Yes	Yes	Yes
Flink	Batch and Streaming	Java, Scala	Yes	Yes	Yes	No

data in real-time fashion. The processing of batch data is treated as a special case in Kappa. The diagram specified below gives the architecture of Flink (Fig. 4; Table 1).

## 10 Conclusion

This paper specifies various comparisons of the tools that can be used in big data analytics. According to the comparison chart given above, it is clear that Map Reduce technique is better only for batch processing system, whereas Spark and Flink can work efficiently on batch processing as well as on streaming data. Fault tolerance is provided in all the techniques, but again, Map Reduce does not support in-memory processing and low latency. According to the survey done on various technologies, Spark is the most efficient framework that can give efficient and accurate results.

## References

- Demchenko Y, Grosso P, de Laat C, Membrey P. Addressing big data issues in scientific data infrastructure. In: 2013 international conference on collaboration technologies and systems (CTS), San Diego, 2013. IEEE, pp 48–55
- Cox M, Ellsworth D. Managing big data for scientific visualization. In: ACM Siggraph '97 course #4 exploring giga-byte datasets in real-time: algorithms, data management, and time-critical design, August, 1997
- Bekkerman R, Bilenko M, Langford J (2011) Scaling up machine learning: parallel and distributed approaches. Cambridge University Press, Cambridge

4. Ni Z Comparative evaluation of spark and stratosphere. Thesis, KTH Royal Institute of Technology; 2013
5. Bu Y, Howe B, Balazinska M, Ernst MD (2010) HaLoop: efficient Iterative data processing on large clusters. Proceedings VLDB Endowment 3(1):285–296
6. Jakovits P, Srivama SN (2014) Evaluating MapReduce frameworks for iterative scientific computing applications. In: 2014 International conference on high performance computing & simulation; 2014. pp 226–33
7. Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H, Seth S, Saha B, Curino C, O’Malley O, Radia S, Reed B, Baldeschwieler E. Apache Hadoop YARN: yet another resource negotiator. In: Proceedings of the 4th annual symposium on cloud computing; 2013
8. Fernández A, del Río S, López V, Bawakid A, del Jesus MJ, Benítez JM, Herrera F (2014) Big data with cloud computing: an insight on the computing environment, MapReduce, and programming frameworks. Wiley Interdiscip Rev Data Min Knowl Discov. 4(5):380–409
9. Lin J, Kolcz A. Large-scale machine learning at twitter. In: Proceedings of the 2012 ACM SIGMOD international conference on management of data; 2012. pp 793–804
10. Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. In: Proceedings of the 6th symposium on operating systems design and implementation; 2004
11. Malewicz G, Austern MH, Bik AJC, Dehnert JC, Horn I, Leiser N, and Czajkowski G (2010) Pregel: A system for large-scale graph processing. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data; 2010. pp 135–45
12. Attenberg J (2015) Conjecture: scalable machine learning in Hadoop with Scalding. 2014. <https://codeascraft.com/2014/06/18/conjecture-scalable-machine-learning-in-hadoop-with-scalding/>. Accessed 1 Jun 2015
13. Zaharia M, Chowdhury M, Das T, Dave A (2012) Fast and interactive analytics over Hadoop data with Spark. USENIX Login 37(4):45–51
14. White T (2012) Hadoop: the definitive guide, 3rd edn. O’Reilly Media, Inc., Sebastopol, CA
15. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark (2010) Cluster Computing with Working Sets. In: Proceedings of the 2nd USENIX conference on hot topics in cloud computing
16. Cai Z, Gao J, Luo S, Perez LL, Vagena Z, Jermaine C. A comparison of platforms for implementing and running very large scale machine learning algorithms. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data (SIGMOD’14) 2014, pp 1371–1382
17. Zhang H, Tudor BM, Chen G, Ooi BC (2014) Efficient in-memory data management: an analysis. Proc VLDB Endowment 7(10):6–9
18. Singh J (2014) Big data analytic and mining with machine learning algorithm. Int J Inform Comput Technol 4(1):33–40
19. Ousterhout K, Rasti R, Ratnasamy S, Shenker S, Chun B (2015) Making sense of performance in data analytics frameworks. In: Proceedings of the 12th USENIX symposium. On networked systems design and implementation (NSDI 15)
20. Shahrivari S, Jalili S (2014) Beyond batch processing : towards real-time and streaming big data. Computers 3(4):117–129

# Copy-Move Forgery Detection Methods: A Critique



Monika Kharanghar and Amit Doegar

**Abstract** With the advancement of image editing tools, legitimacy and creditability of the images are put in jeopardy. There are various types of image forgeries techniques like splicing, retouching, false captioning, and copy-move. But the most predominant forgery techniques are splicing and copy-move forgery. In copy-move forgery, a part of the image is copied in the same image and in splicing another image in order to conceal or to duplicate the information residing in the image. In this critique, the copy-move forgery detection (CMFD) methods from 2013 onwards are reviewed under four categories—block-based, keypoint-based, hybrid-, and deep learning-based methods.

**Keywords** CMFD · Block-based methods · Keypoint-based methods · Hybrid methods · Deep learning based methods

## 1 Introduction

Image forgery is the technique of manipulating the information provided in an image either in the same image or in other images. An image says more than a thousand words. In this epoch of multimedia technology where access to pictures is quite effortless, that thousands of words can be manipulated in various ways and can be spread to thousands of kilometers in the blink of an eye. As being inexpensive, handy, and compelling, many image editors like Adobe Photoshop, GIMP, Pixlr, and Corel DRAW have made it easy to employ digital images in a wrongful way. The images are molded to such a degree that they are not vulnerable through the naked eyes questioning their perseverance. It led to a decrease in the tenability of the images that eventually lead to distrust in the integrity of the content in the images.

The image forgery detection techniques are classified into two categories [1, 2]:

1. Active techniques;
2. Passive techniques.

---

M. Kharanghar (✉) · A. Doegar

Computer Science & Engineering Department, NITTTR, Chandigarh 160019, India

e-mail: [monika.kharanghar@gmail.com](mailto:monika.kharanghar@gmail.com)

© Springer Nature Singapore Pte Ltd. 2021

501

V. Goar et al. (eds.), *Advances in Information Communication Technology and Computing*, Lecture Notes in Networks and Systems 135,

[https://doi.org/10.1007/978-981-15-5421-6\\_49](https://doi.org/10.1007/978-981-15-5421-6_49)

The active techniques are mainly consisting of digital signatures and watermarks. The signatures or watermarks are ingrained into the image and later to attest the image authentication; the signatures or watermark can be fetched. If the fetched signature or watermark matches with the original one, then it is said to be safe; otherwise, it is considered as forged. But due to the lack of information about watermark limits the scope of these techniques.

In contrast, passive techniques thoroughly analyze the composition and constitution of an image to detect any kind of forgery. Passive techniques are applied when either there is no information about digital watermark/signature or there is a prior assumption of not having any watermark or signature. Passive forgery detection techniques are the most appropriate techniques to detect tampering accompanied by various post-processing operations. As many post-processing operations bring out many inconsistencies in the image composition, these inconsistencies favor the passive techniques to detect forgery.

Copy-move forgery detection is a passive or blind image tampering detection in which one or more regions have been transcribed and pasted within the same image.

It can be easy to detect the copied part but due to the application of the post-image processing operations and geometric operations make it hard for forgery detection techniques. These image processing operations can be classified into two categories: intermediate processing operations and post-processing operations. Intermediate operations are applied for the structural harmonization and correlation between the copied region and the target image. These operations may consist of rotation, scaling, illumination or chrominance modification, mirroring, etc. In practice of forgery, the intermediate processing can be applied in the sequence of two or more operations. Whereas post-processing operations consist of JPEG compression, blurring or additive noise are used to hide perceivable touches in the image.

The copy-move forgery detection algorithms are generally differentiated as either block-based algorithms or the keypoint-based algorithms [3, 4].

In this critique, copy-move forgery detection (CMFD) methods have been categorized into four types of methods:

1. Block-based methods;
2. Keypoint-based methods;
3. Hybrid methods;
4. Deep learning-based methods.

## 2 Block-Based Methods

### 2.1 *Texture- and Intensity-Based Methods*

The texture is a crucial feature for the image recognition and classification. Texture and intensity generally target the smoothness, coarseness, and regularity in the image. Therefore, image intensity and texture have been extensively exploited for feature

extraction. The features that represent texture and intensity in CMFD [3] are intensity, pattern (like LBP), and color (like CCV).

Li [5] obtained the features from circular blocks employing rotation-invariant uniform local binary patterns (LBP) and matching is done using locally sensitive hashing (LSH). Davarazni et al. [6] proposed the CMFD method where features are extracted using multi-resolution LBP. Muhammad et al. [7] presented the CMFD method where features are elicited from each SPT sub-band by applying LBP. The LBP histograms of all the sub-bands are collected together to pattern a feature vector. Afterward, feature vectors are fed to an SVM classifier. The SVM used the radial basis function kernel to identify the forgery. Uliyan et al. in [8] proposed the CMFD method where the image was segmented using normalized cut segmentation on the basis of the object in the region, then interest points from each segment were localized using the Hessian algorithm and patterns of interest points were analyzed using CSLBP algorithm. Tralic et al. in [9] combined cellular automata (CA) and LBP to elicit feature vectors for CMFD. Lee et al. [10] proposed the CMFD where HOG is applied to segmented overlapping blocks. Ustubioglu et al. [11] proposed method where overlapping blocks were clustered using color moments and features were extracted using color layout descriptor (CLD) (Table 1).

## 2.2 Frequency Transform-Based Methods

Frequency transform-based algorithms are preferred over other block-based algorithms because of their tenacity to noise and distinguishability of translational and rotational components. Yang et al. [14] employed the fast Walsh–Hadamard transform (FWHT) features to detect CMF; afterward, the matching process was improved using the multi-hop jump (MHJ) algorithm. Ranjani et al. in [15] proposed vector value-based approach to detect the CMF where DCT was applied to describe the overlapping or non-overlapping blocks. The square matrix of order  $n$  was constructed by extracting pixel values from the blocks using row reduction and column reduction techniques. Then, the matrix was sorted lexicographically and phase correlation was computed for the blocks corresponding to the rows on the basis of low contrast values. The above process is repeated until the forged image's matrix was formed. The proposed method is efficient in terms of time and cost. Ustubioglu et al. in [16] presented the block-based algorithms with automatic threshold determination where DCT was used to extract the sixteen elements to form the corresponding feature vectors. The correlation between the feature vectors was checked using the color contents of the corresponding blocks. If the similarity was found, the corresponding shift vector was recorded. The threshold value was figured automatically using Benford's generalized law to check the compression of the image. Shah et al. [17] proposed CMFD method where hybridization of DCT and DWT is used for feature extraction. Parveen et al. [18] proposed the DCT-based CMFD. Mahmood et al. [19] presented the stationary wavelet transform (SWT)-based CMFD method which lessened the false detection as well as computational time. Jaiswal et al. [20]

**Table 1** Texture- and intensity-based CMFD methods

References and year	Preprocessing	Feature extraction	Matching method	Post-processing	Performance	Dataset
Davarazni et al. [6] [2013]	RGB to grayscale conversion, circular block division, Wiener filtering	LBP and its types (MLBP)	Lexicographical sorting, k-d tree	RANSAC	Precision >80% against scaling, JPEG compression, blurring	Self-created
Muhammad et al. [7] [2014]	RGB to YCrCb conversion	SPT and LBP	SVM for classification		Accuracy on CASIA ITDE v1.0, CASIA ITDE v1.0 = 94.89%, CASIA ITDE v2.0 = 97.33%, Columbia color DVMM = 96.39%	CASIA ITDE v1.0, CASIA ITDE v2.0, Columbia color DVMM
Uliyan et al. [8] [2015]	N-cut segmentation	CSLBP, Hessian interest points	Euclidean distance	–	TPR = 92% FPR = 8%	MICC-F220
Tralic et al. [9] [2015]	RGB to grayscale conversion, square block division	Cellular automata and LBP	FLANN, Euclidean distance	–	Precision = 61.7 Recall = 5.2 F-measure = 67.4	CoMoFoD
Ustubioglu et al. [11] [2016]	Square block division	Color moments, CLD	Euclidean distance	Morphological operations	Accuracy = 0.94	CoMoFoD
Lee et al. [12] [2015]	RGB to YCrCb conversion, Square block division	Histogram of orientated Gabor magnitude (HOGM)	Euclidean distance	Calculation of contiguous locations (<60)	False detection ratio = 2.8% Correct detection ratio = 98.8%	CoMoFoD, IMD
Malviya et al. [13]	Noise filtering, square block division, 8Z affine transformation	Auto color correlogram (ACC)	Manhattan distance	–	Precision = 95.65% Recall = 91.67% F1 = 93.62%	CoMoFoD

proposed the shift-invariant SWT-based technique where image is converted into YCbCr and Y channel was disintegrated into four components of SWT. The LL component was divided into  $8 \times 8$  blocks from which mean features were taken and inserted into feature vector with block location. Chen et al. [21] proposed the CMFD method based on fractional quaternion cosine transform (FrQCT) and modified PatchMatch algorithm which was robust against scaling, rotation, JPEG compression, and Gaussian noise (Table 2).

### **2.3 Invariant Image Moments-Based Methods**

Image moments are scalar quantities which are applied to represent the image and to seize some attractive property or interpretation. They can be employed to analyze the shape and perceive an object in the image as they are robust against geometric operations. Many image moments are considered on the basis of the arrangement of orthogonal polynomials and probability distribution like Krawtchouk's moment, exponential moment, Zernike moment, and central moment. Imamoglu et al. [24] presented the CMFD method employing Krawtchouk's moment which is quite sturdy against the post-processing operations, especially the Gaussian blurring. Ryu et al. [25] proposed the Zernike moment-based CMFD method which is sturdy against rotation. Wang et al. [26] proposed quaternion exponent moment (QEM)-based CMFD method where E2LSH is used as a matching method. Al-Qershi et al. [27] proposed the Zernike moments-based CMFD method with new matching method which improved the detection accuracy by 40% in contrast to lexicographical sorting-based matching. Chen et al. [28] obtained the features employing fractional quaternion Zernike moments (FrQZM) from the circular blocks and matched them using the modified PatchMatch algorithm (Table 3).

### **2.4 Log-Polar Transform-Based Methods**

Zhong et al. in [29] presented the discrete analytical Fourier–Mellin transform (DAFMT) to detect the CMF. Luo's rule was used to lessen the background information's interference. The proposed algorithm was robust against translation, scaling, rotation, and minor Gaussian noise. Li et al. [30] adopted the polar harmonic transform (PHT) to describe the contents of circular blocks. Wo et al. [31] presented a CMFD method based on multi-radius PCET that can detect the pasted region with large-scale scaling and rotation. Park et al. in [32] proposed the CMFD method employing up-sampled log-polar Fourier (ULPF) which was robust against scaling and rotation. Yuan et al. [33] proposed the robust CMFD method employing the framework of log-polar expansion and phase correlation. Fadl et al. [34] proposed the polar copy-move (PCM) system to detect the forgeries which were sturdy against rotation, scaling, JPEG compression, brightness, and blurring (Table 4).

**Table 2** Frequency transform-based CMFD methods

References and year	Preprocessing	Feature extraction	Matching method	Post-processing	Performance	Dataset
Ustubioglu et al. [16] [2016]	RGB to YCbCr conversion	DCT with zigzag scan	Benford's generalized law	–		CoMoFoD, Kodak, Google search
Shah et al. [17] [2017]	RGB to grayscale conversion	DCT, DWT	K-means, radix sort	–	Precision = 98.13% Recall = 95.45% F1 score = 0.9677	MICC-F220
Mahmood et al. [22] [2018]	RGB to YCbCr conversion	DCT, SWT	Block distance threshold, block similarity threshold	Morphological operation		CoMoFoD, UCID v2
Jaiswal et al. [20] [2018]	RGB to YCbCr conversion	Shift invariant SWT	Lexicographic sorting, shift vector counter	–		CoMoFoD
Chen et al. [21] [2018]	Square block division	FrQCT	PatchMatch algorithm	Morphological operation	F-measure for GRIP:0.9577 FAU:0.9420	FAU, GRIP
Soni et al. [23] [2017]	Square block division	FWHT	Lexicographical sorting	–	TPR > 96% FPR < 10	CoMoFoD

**Table 3** Invariant image moments-based CMFD methods

References and year	Preprocessing	Feature extraction	Matching method	Post-processing	Performance	Dataset
Imanoglu et al. [24] [2013]	Square block division	Krawtchouk's moment	Lexicographical sorting	Morphological operation	Accuracy > 90% FPR < 0.097	Self-created
Ryu et al. [25] [2013]	RGB to grayscale conversion, square block division	Zernike moments	Locality-sensitive hashing	RANSAC	PDA = 99.4% PFP = 9.7% (rotation)	BOSS image database, self-created
Wang et al. [26] [2018]	Gaussian low-pass filter preprocessing circular block division	QEM	Euclidean locality-sensitive hashing	RANSAC	Detection rate at image level = 0.97 Detection rate at pixel level = 0.88	IMD
Al-Qershi et al. [27] [2015]	RGB to grayscale conversion, square block division	Zernike moments	Bucket distribution of blocks	–	Overall accuracy = 48.93%	Self-created
Chen et al. [28] [2018]	Circular block division	FrQZM	Modified PatchMatch algorithm	–	F-measure for GRIP: 0.9533 FAU: 0.9392	FAU, GRIP

**Table 4** Log-polar transform-based CMFD methods

References and year	Preprocessing	Feature extraction	Matching method	Post-processing	Performance	DATA SET
Zhong et al. [29] [2016]	RGB to grayscale conversion	DAFMFT	Lexicographical sorting, Spearman rank correlation coefficients	–		Self-created
Wo et al. [31] [2016]	Circular block division	Multi-radius PCET	Lexicographical sorting	Morphological operation	F1(blur) = 83.2%	Kodak, IMD
Granty et al. [35] [2016]	8 × 8 block division	PCT	Spectral hashing, Hamming distance, Euclidean distance	Morphological operations	Precision = 0.98 Recall = 1 F1 = 0.99	Self-created
Park et al. [32] [2016]	Gaussian low-pass filtering, circular block division	ULPF	Lexicographical sorting, Euclidean distance	–	F1 = 12.02	IMD
Fadl et al. [34] [2017]	RGB to grayscale conversion, square block division	Polar system conversion, FFT	Radix sort	–	Precision = 85.02–99.5%	DVMM

## 2.5 Dimensionality Reduction-Based Algorithms and Other Miscellaneous Algorithms

Singular value decomposition (SVD) extracts algebraic and geometric features of the image using matrix factorization technique. These extracted features have been extensively employed to identify copy-move forgery because of their cohesion and invariance against rotation and scaling. Kashyap et al. [35, 36] proposed CMFD method which uses SVD as feature extractor on wavelet decomposition of the image and cuckoo search for optimal parameters. Dixit et al. [37] proposed blur-invariant CMFD approach in which SVD is used to extract features from LL band elicited using SWT decomposition providing the detection accuracy of 95% and above.

Mahmood et al. [38] proposed the technique utilizes local binary pattern variance (LBPV)-based features that are elicited from the overlapping blocks of low approximation sub-band of SWT (Table 5).

**Table 5** Dimensionality reduction-based and other miscellaneous method for CMFD

References and year	Preprocessing	Feature extraction	Matching method	Post-processing	Performance	Dataset
Kashyap et al. [36] [2017]	Wavelet decomposition, square block division	SVD	Euclidean distance	Parameter estimation using cuckoo search	Precision = 0.96 Recall = 0.92	Internet
Dixit et al. [37] [2017]	RGB to grayscale conversion, SWT decomposition, Square block division	SVD	Euclidean distance	8-connected neighborhood checking, automated threshold fitting	Detection accuracy without blurring = 99%, detection accuracy with blurring = 95%	Self-created
Mahmood et al. [38] [2017]	RGB to grayscale conversion	SWT, LBP	Block distance threshold, Euclidean distance	Morphological operations	Robust against translation, brightness, noise	CoMoFoD, KLTCI
Kuznetsov et al. [39] [2017]	Image intensity reduction, gradient calculation, expansion in orthogonal basis, ALC enhancement, LBP	Structure pattern and 2D Rabin–Karp rolling Hash function	–	–	TP > 0.7 FP < 0.1	Self-created

### 3 Keypoint-Based Methods

Li et al. in [40] proposed the segmentation-based CMFD in two stages. In the first stage, the image was subdivided into more than 100 patches using SLIC algorithm. The SIFT algorithm was used for the keypoints' detection and description. Then, keypoints in different patches are subjected to matching using KNN. The RANSAC was used to remove the noise from the matrix. In the second stage, the EM algorithm was applied to refine the estimated transform matrix iteratively to eliminate false alarm patches. The results were evaluated to check the vitality against scaling, rotation, noise addition, and JPEG compression. Khayyat et al. in [41] proposed the improved dense scale-invariant feature transform (DSIFT) descriptor to detect the copy-move forgery. They also proposed neighborhood clustering to remove false matching. They improved the DSIFT descriptor first, by using the second- and third-order central moments to identify the dominant orientation and second, by taking circular area contrary to square one to lessen the border effects. The proposed method is vigorous against rotation and many other post-processing methods. Shi et al. [42] proposed CMFD method which uses SIFT as feature descriptor, best bin first (BBF) as feature matching and particle swarm optimization (PSO) for optimal parameters. Jin et al. [43] proposed the SIFT-based CMFD where optimized J-linkage algorithm is used to lessen the computational cost of keypoints clustering.

Kumar et al. in [44] have proposed a hybrid approach in which keypoints are detected using SURF algorithm, and then, the description is done by BRISK algorithm. Mishra et al. [45] implemented CMFD with SURF. The notion of hierarchical agglomerative clustering (HAC) is applied to operate grouping on the matched keypoints achieving TPR of 73.6% and FPR 3.64%. Pandey et al. in [46] explained the precision and accuracy of SIFT, SURF, HOG, SIFT-HOG, and SURF-HOG for detecting the copy-move forgery. They stated the methodology in which keypoints are extracted using SIFT, SURF, and hybrid approach and are matched using Euclidean distance in the first step. Edoardo Ardizzone et al. in [47] presented the approach in which instead of blocks or interest points, the triangles were formed and compared to detect forgery. A Delaunay triangulation was built using the interest points detected by SURF, SIFT, and Harris detectors. The authors stated two methods to represent the objects in the image. In the first method, the image was segmented into the triangles using the dominant colors' extraction and quantitation. The triangle areas and their respective inner angles (in the anticlockwise direction) were also calculated. The similarities between triangles were checked by using sum of absolute deviation (SAD) of the color vectors and of the angles. In the second method, the triangles were formed using SURF and SIFT detectors only and the mean vertex descriptor (MVD) was calculated for each triangle to sort and compare them. The outliers were detected using RANSAC in both methods. Vaishnavi et al. in [48] presented the keypoint-based approach using contrast context histogram (CCH) features. CCH features are extracted from detected keypoints. The outliers were removed using the RANSAC algorithm. The forged regions were localized using the disparity map. Yang et al. in [49] presented the keypoint-based approach using SIFT and KAZE. The features

were extracted and described using SIFT points with 128 feature vectors and KAZE points with 64 feature vectors. The outliers were removed using the image segmentation with SLIC algorithm. The RANSAC algorithm was used to improve the accuracy. Li et al. [50] proposed the CMFD based on SIFT method. Prior to using SIFT, they lowered the contrast threshold of the image and resize it. The hierarchical feature matching is used to match the copied regions. Soni et al. [51] proposed SIFT-based CMFD method where generalized 2NN is used as matching procedure and density-based spatial clustering of application (DBSCAN) as outliers remover. The approach gained TPR > 98 and FPR < 6.8. Raj et al. [52] proposed CMFD method based where image was segmented using SLIC and SURF keypoints are extracted from these segmented patches. Matching procedure consisted of nearest neighbors with affine transformation and expectation maximization algorithm. Isaac et al. in [53] presented keypoints-based approach using HOG features. The keypoints were extracted using Harris Corner points. The features around those keypoints were extracted using the HOG descriptor. Sum of squared differences (SSD) and nearest neighbor distance ratio were used for matching of keypoints. The outliers were removed using the RANSAC algorithm. The results were evaluated at the image level as well as at the pixel level. Muzaffer et al. [54] proposed CMFD method where SURF was utilized as feature descriptor and PSO for optimal parameters. Yeap et al. [57] proposed the CMFD method employing oriented features from accelerated segment test and rotated binary robust independent elementary features (oriented FAST and rotated BRIEF) as the feature extraction algorithm and 2 nearest neighbor (2NN) with hierarchical agglomerative clustering (HAC) as the feature matching algorithm. Niyishaka et al. [56] proposed the CMFD method employing DoG and ORB features (Table 6).

## 4 Hybrid Methods

Manu et al. [58] proposed a method to detect CMF in images employing an over complete segmentation (SLIC) and SURF keypoints. Pun et al. [59] presented the Adaptive over-segmentation and SIFT-based method which is robust against scaling, rotation, JPEG compression, and down-sampling. Zandi et al. proposed the CMFD method where interest points are detected using local density and described employing PCT. A new filtering algorithm was also proposed to reduce false matching pairs. Hashmi et al. in [64] proposed the combination of DyWT and SIFT algorithms for CMFD. Firstly, they divided the image into four parts LL, LH, HL, HH employing DyWT and afterward applied the SIFT algorithm on the LL part to extract the features as LL part enclosed more information than any other parts. The extracted features are subjected to matching to localize the forged region. J. Zheng et al. in [60] proposed the fusion of block-based and keypoints-based approaches. Firstly, the image was segmented into non-imbricated regions using SLIC algorithm. Then for the keypoints detection, SIFT was used and for matching the keypoints g2NN strategy is used. The outliers and false alarms were removed using RANSAC and hierarchical clustering.

**Table 6** Keypoint-based CMFD methods

References and year	Preprocessing	Feature extraction	Matching method	Performance	Dataset
Li et al. [40] [2015]	Image segmentation	SIFT	k-d sort, KNN search, EM algorithm	Precision = 0.86 Recall = 0.88, F1 values = 0.87	IMD, MICC-F600
Shi et al. [42] [2016]	RGB to grayscale conversion	SIFT	Best Bin First (BBF)	Precision = 0.99	IMD
Kumar et al. [44] [2015]	RGB to grayscale conversion	SURF and BRISK	kNN search, Euclidean distance	–	CoMoFoD, IMD
Mishra et al. [45] [2013]	RGB to grayscale conversion	SURF	Euclidean distance, HAC	TPR = 73.6% FPR = 3.64%	MICC-F220
Edoardo Ardizzone et al. [47] [2015]	–	SIFT, SURF, Harris, Delaunay triangulation	Triangle matching using SAD of color vectors and the angles, mean vertex descriptor (MVD)	Link precision = 67.4 Precision = 61.7 Recall = 5.2	CVIP (CMF)
Vaishnavi et al. [48] [2015]	–	CCH	k-means, Euclidean distance	TPR = 78.55% FPR = 35%	MICC-220
Yang et al. [49] [2017]	–	SIFT, KAZE	Euclidean Distance, $n$ -best NN	Recall = 87.92% Precision = 97.08%	IMD
Li et al. [50] [2018]	Lowering of contrast threshold and resizing of image	SIFT	Hierarchical feature point matching	F-image = 98.97% F-pixel = 94.28%	FAU, GRIP, MICC-F200, MICC-F600, coverage, CMH
Issac et al. [53] [2016]	RGB to grayscale conversion	Harris corners points and HOG	Sum of squared differences (SSD), nearest neighbor distance ratio	Precision > 0.9	CVIP, CoMoFoD

(continued)

**Table 6** (continued)

References and year	Preprocessing	Feature extraction	Matching method	Performance	Dataset
Pandey et al. [55] [2015]	–	SIFT(robust), SURF (fast)	g2NN matching	Accuracy (robust) = 100% Accuracy (fast) = 98.5%	MICC-F220
Niyishaka et al. [56] [2018]	Sobel edge detection	DoG and ORB	Hamming distance	Precision = 96.47 Recall = 91.33 F1 = 93.82	MICC-F220, MICC-F8, CoMoFoD
Yeap et al. [57] [2018]	RGB to grayscale conversion, image resizing	ORB	2NN, HAC	Accuracy (F600) = 84.33% Accuracy (F2000) = 82.74%	MICC-F600, MICC-F2000

The forgery in smooth regions is detected using Zernike moments followed by lexicographic sorting of feature vectors. Because of the sparse nature of keypoints, SLIC algorithm was applied again for the segmentation of the image. The regions with matching points were considered forged. The proposed method was evaluated at the image level as well as at the pixel level (Table 7).

## 5 Deep Learning Methods

Rao et al. in [65] stated automated hierarchical feature representations learning model to detect splicing and copy-move forgeries. They proposed the CNN model with eight convolutional layers and a fully connected layer with a 2-way classifier. In the first convolutional layer, the kernel weights were set with the 30 basic high-pass filters to boost the generalization ability and expedite the network's convergence. The model was pre-trained with RGB color images. Zhang et al. in [66] proposed CNN-based models to detect copy-move forgery. In the fundamental models with two forms, Siamese and pseudo-Siamese, there were three convolutional layers with two max-pooling layers and two fully connected layers with a softmax layer. The input to the model was the image pair, copied one (C) and original one (O). Zhang et al. in [67] presented the two-stage deep learning approach for the detection of forged images. Firstly, the image was transformed into the YCrCb space followed by segmentation into  $32 \times 32$  patches. Each patch was subjected to the three-level 2D Daubechies wavelet decomposition to obtain the complex features. The patches were inputted to the SAE model for learning those complex features followed by multilayer perceptron (MLP) layer. Then, after SAE processing, the

**Table 7** Hybrid methods For CMFD

References and year	Preprocessing	Feature extraction	Matching method	Post-processing	Performance	Dataset
Pun et al. [59] [2015]	Segmentation into irregular blocks,	SIFT	Correlation Coefficients map, Euclidean distance	Morphological operations	Image level precision: 96%, Recall: 100% F1: 97.96% Pixel Level precision: 97.22% Recall: 83.73% F1: 89.97%	IMD
Zheng et al. [60] [2016]	Segmentation using SLIC	SIFT, Zernike moments	g2NN, Euclidean distance	Morphological operations	Precision = 0.8851 Recall = 0.8648 F1-measure = 0.8717	CoMoFoD, MICC-F220, Internet
Soni et al. [61] [2018]	RGB to grayscale conversion, square block division	SURF and FAST	2NN	–	–	MICC-F600, MICC-F2000, MICC-F8
Das et al. [62] [2018]	RGB to grayscale conversion, SWT decomposition	SIFT	Euclidean distance, agglomerative hierarchical clustering	RANSAC	Accuracy = 93% FPR = 4% FNR = 10%	MICC-F220
Zandi et al. [63] [2016]	Interest point detection using local density	PCT	Adaptive matching	Estimate transformation using RANSAC, polar decomposition	–	IMD, SBU-CM16

contextual information was integrated from the patches to find the tampering of the image. Zhou et al. in [68] presented rich model CNN (rCNN) with blocking strategy for image forgery detection. After the CNN processing, the classification of the input image based on the feature vectors is done by the SVM classifier. Zhou et al. [69] proposed the algorithm which extracts the color moments, color layout descriptors (CLD), color and edge directivity descriptor(CEDD), fuzzy color and texture histogram(FCTH), scalable color descriptor, and edge histogram descriptor (EHD) from image blocks. These elicited features are fed to compositional pattern-producing network (CPPNs) for forgery classification. Liu et al. [70] proposed the CMFD method based on the convolutional kernel network (CKN). The method involved the adaptive segmentation by convolutional-oriented boundaries (COB), keypoints detection by segmentation-based keypoint distribution strategy (SKPD), CKN-based feature extraction from keypoints, kNN search, and EM-based algorithm for matching. Wu et al. [71] proposed the two branches deep neural network-based CMFD model, called BusterNet. The two branches, Mani-Det and Simi-Det, are used to detect manipulated regions and cloned regions, respectively. Ouyang et al. [79] proposed the pre-trained ImageNet-based CMFD method. Bayar et al. [80] proposed the CNN-based CMFD method where the constrained convolutional layer is introduced to determine prediction error fields. They presented the CNN architecture named MISLnet with five convolutional layers with batch normalization layer and Tanh activation function for feature extraction and three fully connected layers with Tanh activation function and softmax layer for classification. They constructed the ten dataset each with 60,000 grayscale images using various editing parameters like blurring, median filtering, and noise. The authors have achieved the 99.97% accuracy (Table 8).

## 6 Datasets

The datasets require for the copy-move forgery detection are mainly characterized by the number of forged images, their sizes, types of geometrical, and post-processing attacks and ground truth images (mainly for forgery localization) (Table 9).

## 7 Conclusion

This critique presented the various copy-move forgery detection methods from 2013 onwards. In this critique, unlike other reviews or studies, we have categorized the CMFD methods into four types. Besides the block-based and keypoint-based methods, we have also reviewed the methods based on hybridization or fusion of these two methods as well as deep learning methods. Due to the recent upgrading deep learning methods and tools, it is quite easier to do the forgery detection with minimum small human intervention. But considering the involvement of geometrical

**Table 8** Deep learning-based methods for CMFD

References	Techniques/features	Advantages	Disadvantages	Datasets
Rao et al. [65]	SRM-CNN, Xavier-CNN, patch sampling, SVM classification	Satisfying detection performance	–	CASIA v1.0, CASIA v2.0, Columbia gray DVMM
Zhang et al. [66]	Siamese, pseudo-Siamese, 2-channel, Hybrid 2-channel Siamese	Accuracy rate = 96.99%	More negative sample, small training set, and less deep network	INRIA Copydays, CoMoFoD, Image Manipulation, MICC-F220, MICC-F2000
Zhang et al. [67]	SAE, 450 dimensional 3 Level 2D Daubechies wavelet Decomposition	The forged region is localized Overall accuracy = 91.09%	Large training time consumption	CASIA v1.0, CASIA v2.0, Columbia Image Database
Zhou et al. [68]	Tight blocking, rCNN, SVM classification	Robust against JPEG compression, accuracy rate is over 96.41%	No robustness against other post-processing operations, time-consuming process	CASIA v1.0, CASIA v2.0, Columbia gray DVMM
Zhou et al. [69]	CLD, CEDD, FCTH, scalable color descriptor, EHD, CPP network	High accuracy ratios and low false negative values, robust against post-processing operations such as blurring, white Gaussian noise, gamma correction or JPEG compression	No robustness to scaling, rotation and translation attacks	CoMoFoD
Liu et al. [70]	COB, SKPD, CKN, kNN search, EM-based algorithm	Better time consumption	No robustness against other post-processing operations and geometrical operations	CoMoFoD, MICC-F220
Wu et al. [71]	CNN VGG 16 architecture, sigmoid activation, Percentile pooling	Accuracy for opt-in sample is 78%, robust against various attacks	Overall accuracy is very low	CASIA v2.0, CoMoFoD

**Table 9** List of various copy-move dataset

DATASET [72]	No. of images	Size and format of images	Geometrical attacks	Post-processing operation	No. of ground truth images
MICC-F220 [72]	110 original images 110 tampered images	Various sizes, JPG format	Rotation and scaling	–	–
MICC-F600 [72]	440 Original images 160 tampered images	Various sizes, JPG format	Rotation and scaling	–	160
MICC-F2000 [72]	1300 original images 700 tampered images	2048 × 1536 pixels, JPG format	Rotation and scaling	–	–
MICC-F8multi [72]	8 tampered images	Various sizes, JPG format	–	–	–

(continued)

**Table 9** (continued)

DATASET	No. of images	Size and format of images	Geometrical attacks	Post-processing operation	No. of ground truth images
CoMoFoD [73]	200 image sets (Small), 60 image sets (large)	512 × 512 pixels (small), 3000 × 2000 pixels (large), PNG and JPEG format	translation, rotation, scaling, combination, distortion	Blurring, noise addition, color reduction, brightness change, JPEG compression	400 (small), 120 (large)
COVERAGE [74]	100 original images 100 tampered images	400 × 416pixels, TIF format	Translation, scaling, rotation, free form, combination	Illumination	300
CASIA ITDE v1.0 [75]	800 authentic image 921 tampered images	384 × 256 pixels, JPEG format	Scaling, rotation, distortion and combination	—	—
CASIA ITDE v2.0 [75]	7200 authentic image 5123 tampered image	320 × 240 to 800 × 600 pixels, BMP, TIF, and JPEG format	Scaling, rotation, distortion, and combination	Blurring, JPEG compression	—

(continued)

**Table 9** (continued)

DATASET	No. of images	Size and format of images	Geometrical attacks	Post-processing operation	No. of ground truth images
Image Manipulation Dataset (IMD) [76]	48 image sets	PNG format	Rotation, scaling, and combination	Resampling, Gaussian noise, JPEG compression	48
GRIP [77]	100 original images 100 tampered images	400 × 416 pixels, PNG format	–	–	100
SBUCM16 [78]	240 tampered images	PNG and JPG format	Rotation	Blurring, noise, JPEG compression	240
CVIP (CMF) [47]	50 original images 1060 tampered images	1000 × 700/700 × 1000 pixels, BMP format	Rotation and scaling	–	1060

and post-processing methods, it has become more tedious to detect forgeries. The localization of forged regions is still a great challenge to deep learning methods. The hybrid methods have also overcome the some shortcomings of both block-based and keypoint-based methods but not to the great extent.

## References

- Chauhan D, Kasat D, Jain S, Thakare, V (2016) Survey on keypoint based copy-move forgery detection methods on image. *Procedia—Procedia Comput Sci* 85:206–212
- Warbhe AD, Dharaskar RV, Thakare VM (2016) A survey on keypoint based copy-paste forgery detection techniques. *in Phys Procedia* 78(December 2015):61–67
- Warif NBA et al (2016) Copy-move forgery detection: survey, challenges and future directions. *J Netw Comput Appl* 75:259–278
- Soni B, Das PK, Thounaojam DM (2017) CMFD: a detailed review of block based and key feature based techniques in image copy-move forgery detection. *IET Image Proc* 12(2):167–178
- Li Y (2013) Image copy-move forgery detection based on polar cosine transform and approximate nearest neighbor searching. *Forensic Sci Int* 224(1–3):59–67
- Davarzani R, Yaghmaie K, Mozaffari S, Tapak M (2013) Copy-move forgery detection using multiresolution local binary patterns. *Forensic Sci Int* 231(1–3):61–72
- Muhammad G, Al-Hammadi MH, Hussain M, Bebis G (2014) Image forgery detection using steerable pyramid transform and local binary pattern. *Mach Vis Appl* 25(4):985–995
- Uliyan DM, Jalab HA, Wahab AWA Copy move image forgery detection using Hessian and center symmetric local binary pattern. In: 2015 IEEE conference on open systems (ICOS), 2015, pp 7–11
- Tralic D, Grgic S, Sun X, Rosin PL (2016) Combining cellular automata and local binary patterns for copy-move forgery detection. *Multimed Tools Appl* 75(24):16881–16903
- Lee JC, Chang CP, Chen WK (2015) Detection of copy-move image forgery using histogram of orientated gradients. *Inf Sci* 321:250–262
- Ustubioglu B, Ulutas G, Ulutas M, Nabiiev VV (2016) Improved copy-move forgery detection based on the CLDs and colour moments. *The Imaging Sci J* 64(4):215–225
- Lee JC (2015) Copy-move image forgery detection based on Gabor magnitude. *J Vis Commun Image Represent* 31(July):320–334
- Malviya AV, Ladhae SA (2016) Pixel based image forensic technique for copy-move forgery detection using auto color correlogram. *Procedia Comput Sci* 79:383–390
- Yang B, Sun X, Chen X, Zhang J, Li X (2013) An efficient forensic method for copy-move forgery detection based on DWT-FWHT. *Radioengineering* 22(4):1098–1105
- Ranjani MB(2016) Image duplication copy move forgery detection using discrete cosine transforms method matrix sorting—row wise matrix sorting—colum wise 11(4):2671–2674
- Ustubioglu B, Ulutas G, Ulutas M, Nabiiev VV (2016) A new copy move forgery detection technique with automatic threshold determination. *AEU—Int J Electroni Commun* 70(8):1076–1087
- Shah TJ, Banday, MT Copy-move forgery detection using hybrid transform and K-means clustering technique. In: 2017 3rd international conference on applied and theoretical computing and communication technology (iCATccT), 2017, pp. 79–83
- Parveen A, Khan ZH, Ahmad SN (2019) Block-based copy–move image forgery detection using DCT. *Iran J Comput Sci* 0123456789
- Mahmood T, Mehmood Z, Shah M, Khan Z (2018) An efficient forensic technique for exposing region duplication forgery in digital images. *Appl Intell* 48(7):1791–1801
- Jaiswal AK, Srivastava R (2019) Copy-move forgery detection using shift-invariant SWT and block division mean features. In: Khare A, Tiwary US, Sethi IK, Singh N (eds) vol 524. Springer Singapore, Singapore, pp 289–299

21. Chen B, Yu M, Su Q, Li L (2019) Fractional quaternion cosine transform and its application in color image copy-move forgery detection. *Multimed Tools Appl* 78(7):8057–8073
22. Mahmood T, Mehmood Z, Shah M, Saba T (2018) A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform. *J Vis Commun Image Represent* 53:202–214
23. Soni B, Das PK, Thounaojam DM (2017) Blur Invariant block based copy-move forgery detection technique using FWHT features. *Proc Int Conf Watermarking Image Process—ICWIP 2017*:22–26
24. Imamoglu MB, Ulutas G, Ulutas M Detection of copy-move forgery using Krawtchouk moment. In: 2013 8th international conference on electrical and electronics engineering (ELECO), 2013, pp 311–314
25. Ryu S-J, Kirchner M, Lee M-J, Lee H-K (2013) Rotation invariant localization of duplicated image regions based on Zernike moments. *IEEE Trans Inf Forensics Secur* 8(8):1355–1370
26. Wang X, Liu Y, Xu H, Wang P, Yang H (2018) Robust copy-move forgery detection using quaternion exponent moments. *Pattern Anal Appl* 21(2):451–467
27. Al-Qershi OM, Khoo BE (2015) Enhanced matching method for copy-move forgery detection by means of Zernike moments 485–497
28. Chen B, Yu M, Su Q, Shim HJ, Shi Y-Q (2018) Fractional Quaternion Zernike moments for robust color image copy-move forgery detection. *IEEE Access* 6:56637–56646
29. Zhong J, Gan Y (2016) Detection of copy-move forgery using discrete analytical Fourier-Mellin transform. *Nonlinear Dyn* 84(1):189–202
30. Li L, Li S, Zhu H, Wu X (2014) Detecting copy-move forgery under affine transforms for image forensics. *Comput Electr Eng* 40(6):1951–1962
31. Wo Y, Yang K, Han G, Chen H, Wu W (2016) Copy-move forgery detection based on multi-radius PCET. *IET Image Proc* 11(2):99–108
32. Park C-S, Kim C, Lee J, Kwon G-R (2016) Rotation and scale invariant upsampled log-polar fourier descriptor for copy-move forgery detection. *Multimed Tools Appl* 75(23):16577–16595
33. Yuan Y, Zhang Y, Chen S, Wang H (2017) Robust region duplication detection on log-polar domain using band limitation. *Arab J Sci Eng* 42(2):559–565
34. Fadl SM, Semary NA (2017) Robust copy-move forgery revealing in digital images using polar coordinate system. *Neurocomputing* 265:57–65
35. Grantly REJ, Kousalya G (2016) Spectral-hashing-based image retrieval and copy-move forgery detection. *Aust J Forensic Sci* 48(6):643–658
36. Kashyap A, Agarwal M, Gupta H (2018) Detection of copy-move image forgery using SVD and cuckoo search algorithm. *Int J Eng Technol* 7(2.13):79
37. Dixit R, Naskar R, Mishra S (2017) Blur-invariant copy-move forgery detection technique with improved detection accuracy utilising SWT-SVD. *IET Image Proc* 11(5):301–309
38. Mahmood T, Irtaza A, Mehmood Z, Tariq Mahmood M (2017) Copy-move forgery detection through stationary wavelets and local binary pattern variance for forensic analysis in digital images. *Forensic Sci Int* 279:8–21
39. Kuznetsov A, Myasnikov V (2017) A new copy-move forgery detection algorithm using image preprocessing procedure. *Procedia Eng* 201:436–444
40. Li J, Li X, Yang B, Sun X (2015) Segmentation-based image copy-move forgery detection scheme. *IEEE Trans Inf Forensics Secur* 10(3):507–518
41. Khayyat ARH, Sun X, Rosin PL Improved DSIFT descriptor based copy-rotate-move forgery detection. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol 9431, pp 642–655
42. Wenchang S, Fei Z, Bo Q, Bin L (2016) Improving image copy-move forgery detection with particle swarm optimization techniques. *China Commun* 13(1):139–149
43. Jin G, Wan X (2017) An improved method for SIFT-based copy-move forgery detection using non-maximum value suppression and optimized J-linkage. *Sig Process Image Commun* 57:113–125
44. Kumar S et al (2015) A fast keypoint based hybrid method for copy move forgery detection. *Int J Comput Dig Syst* 4(2):91–99

45. Mishra P, Mishra N, Sharma S, Patel R (2013) Region duplication forgery detection technique based on SURF and HAC. *Sci World J* 2013:1–8
46. Pandey RC, Agrawal R, Singh SK, Shukla KK (2015) Passive copy move forgery detection using SURF, HOG and SIFT features. In: Satapathy SC, Biswal BN, Udgata SK, Mandal JK (eds) *Advances in intelligent systems and computing*, vol 327. Springer International Publishing, Cham, pp 659–666
47. Ardizzone E, Bruno A, Mazzola G (2015) Copy–move forgery detection by matching triangles of keypoints. *IEEE Trans Inf Forensics Secur* 10(10):2084–2094
48. Vaishnavi D, Subashini TS (2015) A passive technique for image forgery detection using contrast context histogram features. *Int J Electron Secur Digit Forensics* 7(3):278
49. Yang F, Li J, Lu W, Weng J (2017) Copy-move forgery detection based on hybrid features. *Eng Appl Artif Intell* 59:73–83
50. Li Y, Zhou J (2019) Fast and effective image copy-move forgery detection via hierarchical feature point matching. *IEEE Trans Inf Forensics Secur* 14(5):1307–1322
51. Soni B, Das PK, Thounaojam DM (2018) Keypoints based enhanced multiple copy-move forgeries detection system using density-based spatial clustering of application with noise clustering algorithm. *IET Image Proc* 12(11):2092–2099
52. Raj R, Joseph N (2016) Keypoint extraction using SURF algorithm for CMFD. *Procedia Comput Sci* 93(September):375–381
53. Isaac MM, Wilscy M (2016) A key point based copy-move forgery detection using HOG features. In: 2016 international conference on circuit, power and computing technologies (ICCPCT), 2016, pp 1–6
54. Muzaffer G, Ulutas G, Gedikli E PSO and SURF based digital image forgery detection. In: 2017 international conference on computer science and engineering (UBMK), 2017, pp 688–692
55. Pandey RC, Singh SK, Shukla KK, Agrawal R (2015) Fast and robust passive copy-move forgery detection using SURF and SIFT image features. In: 9th international conference on industrial and information systems, ICIIS 2014
56. Niyishaka P, Bhagvat C (2018) Digital image forensics technique for copy-move forgery detection using DoG and ORB. In: Bolc L, Tadeusiewicz R, Chmielewski LJ, Wojciechowski K (eds), vol. 7594. Springer, Berlin Heidelberg, pp 472–483
57. Yeap YY, Sheikh UU, Rahman AA-HA (2018) Image forensic for digital image copy move forgery detection. In: 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA), pp 239–244
58. Manu VT, Mehtre BM Detection of copy-move forgery in images using segmentation and SURF. *Adv Intell Syst Comput* 425:645–654
59. Pun C-M, Yuan X-C, Bi X-L (2015) Image forgery detection using adaptive oversegmentation and feature point matching. *IEEE Trans Inf Forensics Secur* 10(8):1705–1716
60. Zheng J, Liu Y, Ren J, Zhu T, Yan Y, Yang H (2016) Fusion of block and keypoints based approaches for effective copy-move image forgery detection. *Multidimens Syst Signal Process* 27(4):989–1005
61. Soni B, Das PK, Meitei Thounaojam D (2018) Improved block-based technique using SURF and FAST keypoints matching for copy-move attack detection. In: 2018 5th international conference on signal processing and integrated networks, SPIN 2018, pp 197–202
62. Das T, Hasan R, Azam MR, Uddin J A robust method for detecting copy-move image forgery using stationary wavelet transform and scale invariant feature transform. In: 2018 international conference on computer, communication, chemical, material and electronic engineering (IC4ME2), 2018, pp 1–4
63. Zandi M, Mahmoudi-Aznaveh A, Talebpour A (2016) Iterative copy-move forgery detection based on a new interest point detector. *IEEE Trans Inf Forensics Secur* 11(11):2499–2512
64. Hashmi MF, Anand V, Keskar AG (2014) Copy-move image forgery detection using an efficient and robust method combining un-decimated wavelet transform and scale invariant feature transform. *AASRI Procedia* 9:84–91
65. Rao Y, Ni J (2017) A deep learning approach to detection of splicing and copy-move forgeries in images. In: 8th IEEE international workshop on information forensics and security, WIFS 2016

66. Zhang J, Zhu W, Li B, Hu W, Yang J (2016) Image copy detection based on convolutional neural networks. In: CCIS, vol 2, pp 111–121
67. Zhang Y, Goh J, Win LL, Thing V (2016) Image region forgery detection: a deep learning approach. *Cryptol Inf Secur Ser* 14:1–11
68. J. Zhou, J. Ni, and Y. Rao (2017) Block-based convolutional neural network for image forgery detection vol 10431, Kraetzer C, Shi Y-Q, Dittmann J, Kim HJ (eds) Springer International Publishing, Cham, pp 65–76
69. Zhou H, Shen Y, Zhu X, Liu B, Fu Z, Fan N (2016) Digital image modification detection using color information and its histograms. *Forensic Sci Int* 266:379–388
70. Liu Y, Guan Q, Zhao X (2018) Copy-move forgery detection based on convolutional kernel network. *Multimed Tools Appl* 77(14):18269–18293
71. Wu Y, Abd-Almageed W, Natarajan P (2018) BusterNet: detecting copy-move image forgery with source/target localization, pp 170–186
72. Amerini I, Ballan L, Caldelli R, Del Bimbo A, Serra G (2011) A SIFT-based forensic method for copy-move attack detection and transformation recovery. *IEEE Trans Inf Forensics Secur* 6(3):1099–1110
73. Tralic D, Zupancic I, Grgic S, Grgic M (2013) CoMoFoD—new database for copy-move forgery detection. In: Proceedings of 55th international symposium ELMAR-2013 (September):25–27
74. Wen B, Zhu Y, Subramanian R, Ng T-T, Shen X, Winkler S (2016) COVERAGE—a novel database for copy-move forgery detection. In: 2016 IEEE international conference on image processing (ICIP), pp 161–165
75. Dong J, Wang W, Tan T (2013) CASIA image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing, pp 422–426
76. Christlein V, Riess C, Jordan J, Riess C, Angelopoulou E (2012) An evaluation of popular copy-move forgery detection approaches. *IEEE Trans Inf Forensics Secur* 7(6):1841–1854
77. Cozzolino D, Poggi G, Verdoliva L (2015) Efficient dense-field copy-move forgery detection. *IEEE Trans Inf Forensics Secur* 10(11):2284–2297
78. Zandi M, Mahmoudi-Aznaveh A, Mansouri A (2014) Adaptive matching for copy-move Forgery detection. In: 2014 IEEE international workshop on information forensics and security (WIFS), 2014, pp 119–124
79. Ouyang J, Liu Y, Liao M (2017) Copy-move forgery detection based on deep learning. In: 2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI), pp 1–5
80. Bayar B, Stamm MC (2018) Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection. *IEEE Trans Inf Forensics Secur* 13(11):2691–2706

# Improving Website by Analysis of Web Server Logs Using Web Mining Tools



Neeraj Kandpal, Devesh Kumar Bandil, and M. S. Shekhawat

**Abstract** Web log data obtained from the server of the website is the key source of enormous hidden information. This information can be obtained from the analysis of web log data. Analysis of web log data is very useful for the management and improvement of the website. It also plays a vital role in the security measures of a website. Analysis of web log data is also very important for analysis of user's behavior and for the maintenance purpose of website. This analysis is possible by many available web usage mining tools. In this paper, we have used Web Log Expert tool for analysis. Web log Expert gives lots of reports and graphs to gain insight into web log data.

**Keywords** Web usage mining · Web log data · Web Log Expert tool · Preprocessing · Pattern discovery

## 1 Introduction

Web log mining is gaining importance due to its ability to show obscure knowledge. Web log data is the record of all types of activities of visitors accessing the website. Analysis of this data is very useful for improving the features of the website and better management of the website. Web usage mining can be divided into three parts. Web content mining covers the mining of webpage for its contents like text, image, and table. Web structure mining gives insight into the different hyperlinks structure of the website. Web usage mining produces behavior of visitors while traversing the website. It gives the various patterns, which are representing knowledge of the user's

---

N. Kandpal (✉) · D. K. Bandil  
Suresh Gyan Vihar University, Jaipur, Rajasthan, India  
e-mail: [neerajkandpal11@gmail.com](mailto:neerajkandpal11@gmail.com)

M. S. Shekhawat  
Department of Physics, Government Engineering College, Bikaner, Rajasthan, India

attitude and habits at the time of visiting the website. Again web log mining or web usage mining comprises of preprocessing, pattern discovery, and pattern analysis. Unnecessary data removed in the preprocessing step. Various visible patterns formed in pattern discovery. This pattern gives a set of rules for a specific event based on web log data of user traversal.

## 2 Related Work

Vinod Kumar et al. summarized a wide range of different web mining tools available free or proprietary [1]. Researchers explained the processes involved in web usage mining and summarize important web usage mining tools [2]. ANANDAN BELLIE used WEKA web mining tool for analysis of data of university students [3]. He used k-means algorithm for grouping users of similar behavior using their browser history. Navin Kumar Tyagi et al. used Web Log Expert web mining tool for the study of web log data obtained from sync software [4]. He gave a summary of general statistics and different type of errors. According to him, these error reports are very useful for the web administrator for accessing corrupted and broken links. Similarly, Jigar H. Jobanputra et al. used Web Log Expert tool for improving features of the website. They have given different reports about browsers, visitors, and various useful statistics for web administrators [5]. Ranjena Sriram et al. proposed a preprocessing algorithm by using the concept of hashing key [6]. Neeraj kandpal et al. used web usage mining to improve the efficiency of the website and also for website management [7]. According to researchers pattern, discovery and pattern analysis of web log data are very useful for website administrators and designers [8, 9].

Navjot Kaur et al. used web log mining tool to obtain the best time and best place at the website for placing the advertisement [10]. Researchers defined processes involved in web usage mining [11]. According to them, web usage mining is an effective source for the improvement of the website and makes it user-friendly. Mitali Srivastava et al. were given a survey for different preprocessing methods in web usage mining [12]. Researchers proposed a session identification algorithm where user behavior recorded in matrix format [13]. Nanhay Singh et al. presented a comparative analysis of web log data using the technique of pattern recognition [14]. S. Bhuvaneswari et al. presented a comparative analysis of different web mining tools for the analysis of customer behavior [15].

## 3 Source of Data for Web Usage Mining

Web log data for current research obtained from website ‘rajeducon.com’, which is a departmental website of secondary education, Rajasthan, India. We have taken the data of one month, i.e., June-2019 for analysis and generation of patterns by using web mining tools.

## 4 Web Usage Mining Tools

There are lots of tools available for analysis of web log data. Researchers gave a detailed comparison of different web usage mining software [15]. We have used web usage mining software Web Log Expert for analysis and generating patterns for web log data.

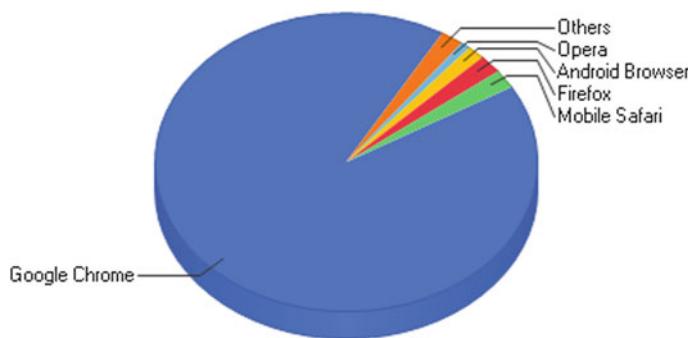
## 5 Results and Discussions

Web log data is analyzed by using Web Log Expert tool. Some of the important results obtained are summarized below.

1. According to general statistics by Web Log Expert, the total number of 690,622 hits encountered by 48,807 visitors and used 262.73 GB bandwidth to reach the total 136,543 pages of the website. The website also found 6683 spider hits, which used 166.34 MB bandwidth.
2. It is found that most of the users used mobile to access the website 'raje-ducon.com'. It prompted us to modify the code according to the need of mobile users to make the website more mobile user-friendly.
3. Most used browsers count is shown in Fig. 1. According to this, the most used

**Table 1** Most used device type

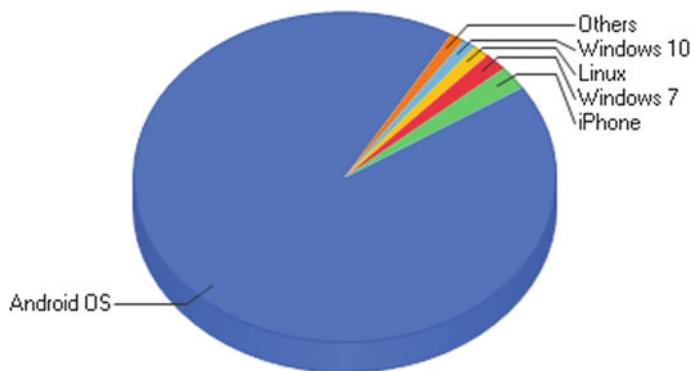
S. No.	Device type	Hits	Visitors	% of total visitors
1.	Mobile	639,896	45,924	94.10
2.	Desktop	42,317	2452	5.02
3.	Tablet	1726	431	0.88
	Total	683,939	48,807	100.00



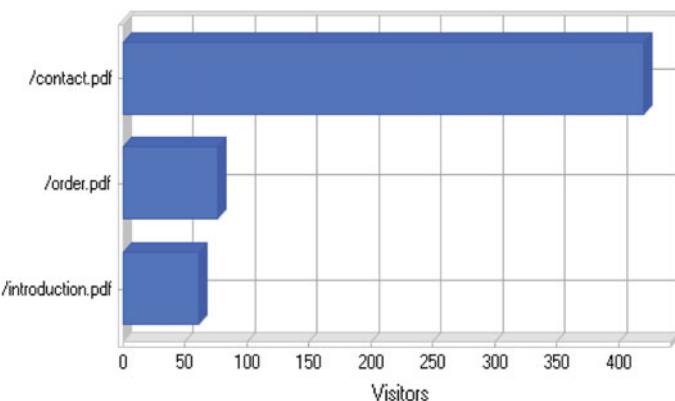
**Fig. 1** Most used browsers

browser for searching the website is Google Chrome. Other important browsers are Opera, Android Browser, Firefox, Mobile Safari, etc.

4. Most used operating systems count is shown in Fig. 2. According to this, Android operating system grabs most of used operating systems in devices accessing 'rajeducon.com'. Some other used operating systems are Windows 10, Linux, Windows 7, and iPhone.
5. The most downloaded files are given in Fig. 3. The contact information of the department is mostly downloaded by the visitors of the website. Introduction and orders pdf files also take some part of downloaded files.
6. The most occurred errors are 404 and 401 as in Table 2. We have checked the web pages according to this report and corrected the erroneous one. We also formed an information page for error 404, which gives the information about the possible cause of the error.
7. Table 3 gives the count of file type mostly downloaded by the visitors of the



**Fig. 2** Most used operating system



**Fig. 3** Most downloaded files

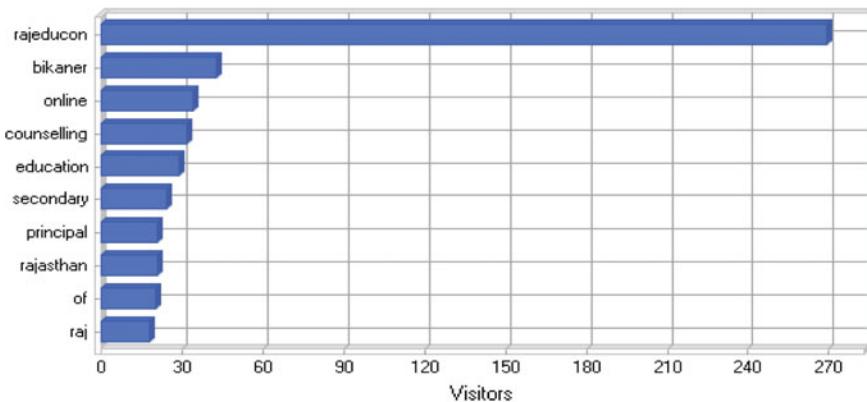
**Table 2** Error encountered

S. No.	Error type	Hits
1	404 Not Found	292,777
2	401 Unauthorized	16

**Table 3** Most requested file types

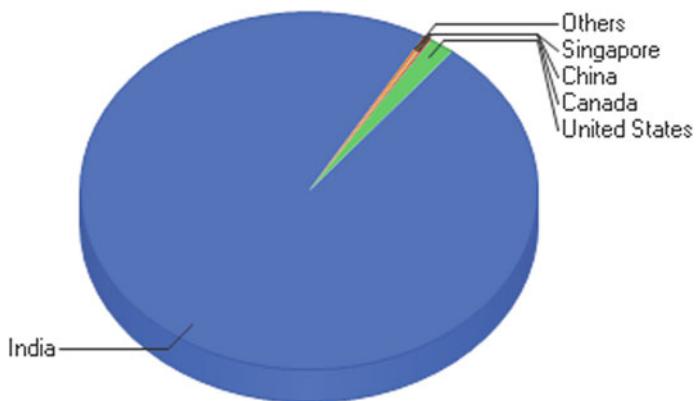
S. No.	File type	Hits	Incomplete request	Bandwidth (KB)
1	Html	102,789	2	250,418
2	Jpg	91,336	40,578	258,210,846
3	CSS	84,481	595	5,363,762
4	Js	56,828	1036	9,150,691
5	Php	33,042	1	108,056
6	Jpeg	22,712	385	1,660,965
7	Pdf	649	33	69,243
8	woff2	1	0	69
	Total	391,838	42,630	274,814,050

- website and bandwidth used. This information is used by researchers [7] to improve the performance of the website. We restricted the use of images and other less useful objects at the peak time of operations. By using these results, we achieved a greater number of user count encountered the website.
8. The keywords are the phrases which used to search website. It is found that rajeducon and Bikaner are the two most used keywords to search online counseling website 'rajeducon.com'.
  9. According to Table 4, Google is the most used search engine by visitors. A very little share taken by other counterparts.

**Fig. 4** Most used keywords

**Table 4** Top-requested search engines

S. No.	Search engine	Visitors
1	Google	13,472
2	Bing	24
3	Yahoo	4
4	DuckDuckGo	1
	Total	13,501

**Fig. 5** Most active countries

10. India is the most active country for website 'rajeducon.com'. Other interested countries are depicted in Fig. 5. Countries other than India coming in Fig. 5 are due to the spiders or bots hits from the servers existing in these countries.

## 6 Conclusions

Web log mining is very crucial for the smooth functioning and security of the website as it brings patterns of user's activity for web administrators. Web Log Expert tool provides the excellent results of web usage mining comparative to other popular tools. It gives a detailed analysis of web log data, which in turn provides effective behavioral patterns. We have found most and least visited pages by users, which show the path for improvement of the structure of the website. Most visited pages are the best places to put important material. Least visited pages checked for possible broken links or irrelevant contents. We have accessed the most popular web browsers and popular operating systems used by visitors. Information about web robots and errors also helps to improve and manage the website. The knowledge of error encountered in accessing the website gives the way to correct broken links of the website. All errors encountered corrected to provide smooth functioning for visitor's navigation.

## References

1. Kumar V, Thakur RS (2017) A brief investigation on web usage mining tools (WUM), Saudi J Eng Technol
2. Kandpal N, Sinha RR, Shekhawat MS (2017) A survey on web usage mining: process, application and tools. Suresh Gyan Vihar Univ J Eng Technol 3(1):19–25
3. Bellie A (2015) Web usage analysis of university students to improve the quality of Internet service. Int J Adv Res Comput Eng Technol (IJARCET), 4(5)
4. Tyagi NK, Solanki AK, Wadhwa M (2010) Analysis of server log by web usage mining for website improvement. Int J Comput Sci Issues 7(4)
5. Jobanputra JH, Soni BD (2016) Enhancing the efficiency of the website by mining web server log. Int J Sci Res Dev 3(12)
6. Sriram R, Mallika R (2016) Innovative pre-processing technique and efficient unique user identification algorithm for web usage mining. Int J Adv Res Comput Sci Softw Eng 6(2)
7. Kandpal N, Singh HP, Shekhawat MS (2019) Application of web usage mining for administration and improvement of online counseling website. Int J Appl Eng Res 14(7):1431–1437
8. Sharma AK, Gupta PC (2013) Analysis of web server log files to increase the effectiveness of the website using web mining tool. Int J Adv Comput Math Sci 4(1):1–8
9. Sharma S, Rai M (2017) Customer behavior analysis using web usage mining. Int J Sci Res Comput Sci Eng 5(6):47–50
10. Kaur N, Aggarwal H (2015) Web log analysis for identifying the number of visitors and their behavior to enhance the accessibility and usability of website. Int J Comput Appl 110(4). 0975–8887
11. Kandpal N, Sinha RR, Shekhawat MS (2016) A study of processes involved in web usage mining. Int J Allied Pract, Res Rev III(XII):01–05
12. Srivastava M, Garg R, Mishra PK (2014) Preprocessing techniques in web usage mining: a survey. Int J Comput Appl 97(18)
13. Chitraa V, Thanamani AS (2011) A novel technique for sessions identification in web usage mining preprocessing. Int J Comput Appl 34(9)
14. Singh N, Jain A, Raw RS (2013) Comparison analysis of web usage mining using pattern recognition techniques. Int J Data Min Knowl Process (IJDKP) 3(4)
15. Bhuvaneswari S, Anand T (2015) A comparative study of different log analyzer tools to analyze user behaviors. Int J Recent Innov Trends Comput Commun 3(5):2997–3002

# A New Approach for Paddy Leaf Blast Disease Prediction Using Logistic Regression



Sree Charitha Kodaty and Balaji Halavath

**Abstract** Paddy is a major agricultural crop. But the production of paddy is hindered by various kinds of diseases. Some of those diseases are leaf blast disease, brown spot disease, bacterial blight disease, etc. Amidst of all these diseases influencing the paddy production, leaf blast disease had a great influence and it is the most destructive diseases that are effecting on paddy crop. Leaf blast is risen by the fungus *Magnaporthe oryzae*. It will affect all the above-ground parts of a paddy crop: leaf, collar, node, neck, parts of panicle, and sometimes leaf sheath. Thus, examining and accurate forecasting for the development of blast disease are significant and early forecasting of the disease is very beneficial. Many former blast disease prediction models were only considering the attribute values but not their correlations. In this paper, logistic regression algorithm is applied for forecasting the occurrence of leaf blast disease for Adilabad district of Telangana state in India during 2007–2017 in order to prevent the paddy fields from disease. With the help of the correlation mining and clustering process among the attributes, we are classifying the attribute sets based on their impact on disease occurrence. Finally, the logistic regression algorithm calculates the leaf blast disease occurrence probability.

**Keywords** Machine learning · Paddy leaf blast disease · Logistic regression

## 1 Introduction

Paddy is a major crop and is also a staple food in various countries and the production of paddy plays an indispensable part of food security in India [1]. Although many diseases are affecting the paddy field, leaf blast disease is the prominent and its impact is very severe on crop yield. Blast, a major disease of rice is affected by the

---

S. C. Kodaty (✉) · B. Halavath

Department of Computer Science and Engineering,

Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, Telangana, India

e-mail: [sreecharithakodaty@gmail.com](mailto:sreecharithakodaty@gmail.com)

B. Halavath

e-mail: [balajimitk@gmail.com](mailto:balajimitk@gmail.com)

fungus *Magnaporthe oryzae*, is widely distributed pathogens of rice, being initiated in nearly all paddy-growing surroundings [2]. Blast disease residue is a significant problem in temperate and sub-tropical paddy production regions, at high elevation in the tropics, and within tropical upland rice.

Disease prediction is an essential part in a disease management system which assists the system to make decisions and comes up with the symptoms when the disease is probably going to arise and when it is going censorious. The prediction of crop ailment, based on meteorological conditions needs concurrent investigation of large amount of meteorological conditions and their composite association with the disease for notable period time (min 10–15 years) [3]. In particular, when the meteorological factors are favorable for the development of blast disease fungus during harvest season, there will be high yield loss. The difficulty of various crop disease process and the vulnerability on some factors are alike that our perception, and thus the prediction expertise of numerous arithmetical approaches is essentially restricted. Besides, there is a vast gap of awareness of the mathematical relations linking the meteorological factors and the particular point of the disease infection cycle [4]. Early prediction and severity detection methods on leaf blast disease help to prevent the paddy fields from blast disease occurrence.

## 2 Literature Review

In 2006 Rakesh Kaundal, Amar S Kapoor, and Gajendra PS Raghava have proposed SVM prediction model to build atmospheric conditions-based forecasting models for crop diseases. Six important environmental variables were taken as predictor variables. Two series models, i.e., cross-location and cross-year were developed and validation is done using fivefold cross-validation procedure. They had developed an SVM-based Web server for rice blast prediction which will be helpful to the plant science community and also the farmers in order to make decisions. Our investigation reveals that SVM is better than existing ml procedures and traditional REG approaches in prediction of crop diseases [5].

In 2011 Shafaulla, Muhammad Aslam Khan, Nasir Ahmed Khan, and Yasir Mahmood have dealt with the effect of meteorological factors for the occurrence and extremity of rice blast during the growing season of 2008. The temperature and occurrence of the blast disease were correlated negatively this shows that the disease will occur when the temperature decreases. Humidity was positively correlated with the disease, and this shows that whenever the humidity increases the occurrence of the disease will also increase and the rainfall is also positively correlated. But more epidemiological studies are required to accurately forecast the disease so that it helps out to minimize the yield loss caused by the blast disease [6].

In 2017 Rini Pal, Dipankar Mandal, and Bhima Sen Naik have conducted a field investigation to discover the impact of meteorological conditions on the extremity of leaf blast disease during 2013–2014 and 2014–2015. It showed that humidness and precipitation were emphatically associated with blast seriousness. On the other

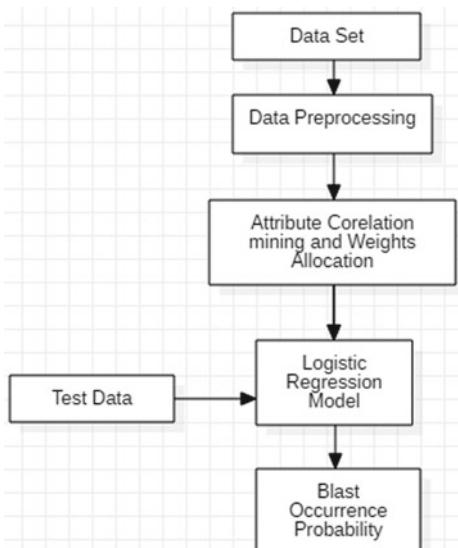
hand, the temperature and disease seriousness were pessimistically associated and show that the disease will increase with the reduction in temperature. The favorable conditions for disease development and spread were high humidity of 90–95% and entire precipitation more than 280 mm along with 28 °C temperature. But still more epidemiological can be included for accurate forecasting of leaf blast disease occurrence [7].

In 2017, S. Bregaglio, P. Titone, and L. Hossard have proposed a model to reduce the occurrence of blast disease in Northern Italy to reduce the economic losses suffered by the farmers. This study has proved that choosing the rice variety is a good procedure to stop the development of disease and its impact on the crop and this study also prove that reduced nitrogen fertilization is successful in reducing the impact of the disease. In this study, three-year investigation (2013–2015) was presented, in which blast intensity was illustrated on four variants which are grown with two nitrogen portions. The progress of disease on leaf and panicle blast is examined via *F*-test for the site, nitrogen usage, rice variant, and year. The regions under disease advancement curves were associated with yield losses via linear regression [8].

### 3 Methodology

See Fig. 1.

**Fig. 1** System architecture



### **3.1 Data Preprocessing**

In this data preprocessing step, the rice blast disease data collected from various sources is streamlined and preprocessed for disease prediction. Duplicate record elimination, missing values imputation, and transformation of data into relational structures are the main aspects of this step.

### **3.2 Attribute Correlation Mining**

In this step, for the input data record attributes, relevant weight values are calculated based on their impact on the crop. Along with this, the attribute correlation mining is calculated to perform the associative mining process. Finally, these attributes will be created as a set.

### **3.3 Logistic Regression for Classification**

Paddy blast disease forecasting system uses logistic regression approach to build the training model. It is a statistical investigation procedure that can be used in forecasting the data value based on the previous monitoring on the data set. The logistic regression model forecasts the dependent data variable by examining the association between one or more independent variables. It is one of the powerful tools for prophecy, which also be used for classifying and forecasting the occurrence of blast disease based on the historical data, and it predicts the likelihood of occurrence of a binary event utilizing a logit function [9].

$$\log \left[ \frac{Y}{1 - Y} \right] = C + B1X1 + B2X2 + \dots$$

#### **3.3.1 Sigmoid Function**

To map the predicted probabilities, we use the sigmoid function. The sigmoid function maps any real input values into output values between the range of 0 and 1. It is defined as:

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

When the test data is passed it will compute the value based on the variables stored in the model. It determines the likelihood of each category. If  $h(x) \geq 0.5$ , then we can say it is class 1 data point else if  $h(x) < 0.5$  then the data point belongs to class 0 [10].

## 4 Results and Discussions

See Fig. 2.

The front end of the paddy blast disease forecasting system is shown in the above figure (Fig. 3).

After registration is done, the user can upload the data for the prediction of blast disease occurrence probability (Fig. 4).

The user input values are compared with the correlation value list and the disease occurrence probability value is displayed. Hence, the outcome is 0.64; so in terms of probability, there is 64% possibility for the disease occurrence (Figs. 5 and 6).

The values that are entered by the user are compared with the correlation value list and the disease occurrence probability value is displayed. Hence, the outcome is 0.4; so in terms of probability, there is 40% possibility of disease occurrence.

Depending on the rice types span in period from 90 to longer than 150 days and with three main crop phases: (1) the vegetative phase—where germination to beginning of panicle (45–100 days), (2) The reproductive phase—where panicle initiation to flowering (35 days), and (3) the maturity phase—where from flowering to mature grain (30 days) (Fig. 7).

**Fig. 2** Login page for blast disease

The image shows a registration form titled "Register Here". The form consists of several input fields and dropdown menus. At the top right is a "Register Here" button. Below it are four text input fields: "First Name" and "Last Name" side-by-side, followed by "Email ID", "New Password", and "Confirm Password" stacked vertically. Underneath these is a "Date of Birth" label followed by three dropdown menus for "Month", "Day", and "Year". At the bottom are two radio buttons for gender: "Female" and "Male", and a large blue "Register" button.

**Fig. 3** Prediction page with blast disease

### Upload Paddy blast disease record

Enter Record Details Here	
Temp	18
Heridity	0.85
Rain_since	2
Rain_till	2
N_use	0.8
Seed_resist	0.6
Herb_pec	0.8
Crp_phse	0.3
Vent_lvl	0.3
<b>Predict Probability</b>	

**Fig. 4** Above tables show the blast disease occurrence probability

Input Dataset Correlation value List									
Temp	Heri dity	Rain since	Rain till	N use	Seed resist	Herb pec	Crp phse	Vent lvl	
0.42	0.7	0.7	0.69	0.42	0.68	0.65	0.38	0.7	

Input Dataset value List									
Temp	Heri dity	Rain since	Rain till	N use	Seed resist	Herb pec	Crp phse	Vent lvl	
0.18	0.85	0.02	0.02	0.8	0.6	0.8	0.3	0.3	

Input Dataset attribute satisfactory List									
Temp	Heri dity	Rain since	Rain till	N use	Seed resist	Herb pec	Crp phse	Vent lvl	
false	false	true	true	false	false	false	true	false	

**Disease Occurrence Probability Value is = 0.64**

**Blast Disease occurred**

The above graph shows when the crop is at the vegetative phase. In this phase, after planting the seeds, the seed germination takes place after 45–100 days during this phase if the rainfall, temperature, nitrogen usage, etc., are at required rate, the chances of occurring the leaf blast is very less, as the paddy leaf blast disease can occur at all growth stages of the crop (Fig. 8).

The above graph shows when the crop is at reproductive phase which is the second phase of the crop. In this phase after panicle initiation took place, the plants

**Fig. 5** Prediction page with no blast disease

### Upload Paddy blast disease record

Enter Record Details Here	
Temp	32
Heridity	0.45
Rain_since	2
Rain_till	1
N_use	0.4
Seed_resist	0.8
Herb_pec	0.5
Crp_phse	0.3
Vent_lvl	0.3
<input type="button" value="Predict Probability"/>	

**Fig. 6** Above table shows the blast disease occurrence probability

Input Dataset Correlation value List									
Temp	Heridity	Rain_since	Rain_till	N_use	Seed_resist	Herb_pec	Crp_phse	Vent_lvl	
0.42	0.7	0.7	0.69	0.42	0.68	0.65	0.38	0.7	

Input Dataset value List									
Temp	Heridity	Rain_since	Rain_till	N_use	Seed_resist	Herb_pec	Crp_phse	Vent_lvl	
0.32	0.45	0.02	0.01	0.4	0.8	0.5	0.3	0.3	

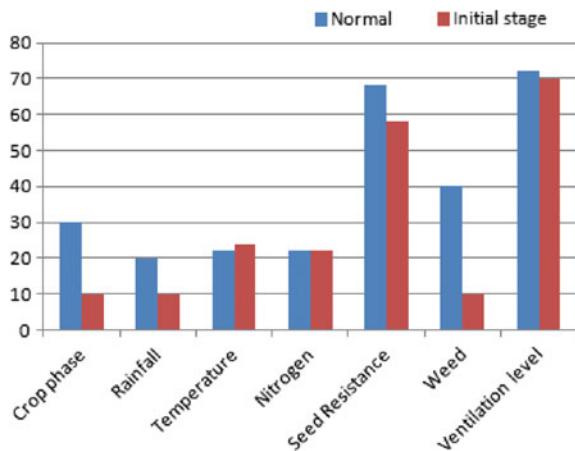
Input Dataset attribute satisfactory List									
Temp	Heridity	Rain_since	Rain_till	N_use	Seed_resist	Herb_pec	Crp_phse	Vent_lvl	
false	true	true	true	true	true	true	false	false	

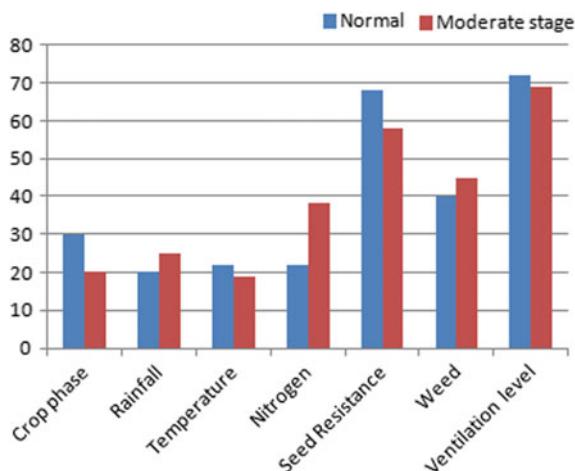
Disease Occurrence Probability Value is = 0.4									
Less Chances to occur Blast Disease									

are transplanted to flowering which takes around 35 days. During this phase when the temperature, rainfall, and nitrogen usage are more than required rate, then the chances of occurrence of the blast disease are slightly more than the previous phase. During the transplantation optimum, crop spacing should be maintained between one plant to another, i.e., there should be proper ventilation if not it becomes overcrowded and the plants must compete with each other for soil nutrients which eventually leads to increased fertilizers usage. And poor air circulation will lead to increase of fungal

**Fig. 7** Above graph shows the disease occurrence at initial stage



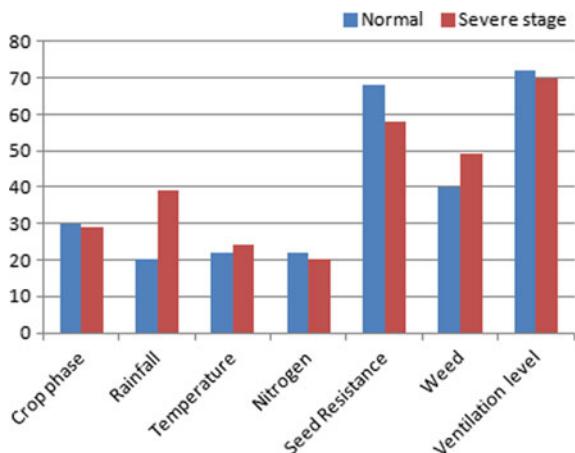
**Fig. 8** Above graph shows the disease occurrence at moderate stage



disease as it can spread easily between the plants if they are spaced so closed. And also during this phase, weed control (unnecessary plants) is important, and the weeds will compete for sunlight, nutrients, and water and it will also reduce the grain quality and it may also attract the pests that may act as a host for the spread of blast disease. (Figure 9).

The above graph shows when the crop is at maturity phase which is the third phase of the crop, this phase is also known as ripening phase, it starts with flowering and ends and it is ready to be harvested, the duration of this phase is 30 days. During this phase, if there are continuous and prolonged periods of rainfall, low temperature during daytime and usage of nitrogen than the required amount can severe the occurrence of blast disease. The number of rainy days in the week can increase the risk of disease occurrence. If the field is affected by the blast disease at

**Fig. 9** Above graph shows the disease occurrence at severe stage



this phase, there will be a huge yield loss. The blast disease is the most destructive disease, and it can kill seedlings or plants up to tillering phase and in later phases, it will reduce the leaf area for grain fill and hence reducing the grain yield. So predicting the occurrence of blast disease can prevent the field from agricultural loss.

## 5 Conclusion

This research aims to develop a user-friendly system for predicting the blast disease occurrence probability with the help of correlation mining and clustering among the attributes. The proposed system is efficient when compared to the other existing blast disease forecasting models. Nevertheless, still additional environmental investigations are needed to increase the prediction process of the disease which have a tendency to eventually reduce the yield losses affected by the disease.

## References

1. Pinki FT, Khatun N, Islam SMM Content based paddy leaf disease recognition and remedy prediction using support vector machine. In: 2017 20th international conference of computer and information technology (ICCIT), Dhaka, 2017, pp 1–5
2. Kim Y, Roh JH, Kim HY (2018) Early forecasting of rice blast disease using long short-term memory recurrent neural networks. Sustainability 10(1):34
3. Bhagawati S, Bhagawati R, Singh K, Nongthombam AKK, Sarmah R, Bhagawati G (2015) Artificial neural network assisted weather based plant disease forecasting system. Int J Recent Innov Trends Comput Commun 3(6):4168–4173
4. Katsantonis D, Kadoglou K, Dramalis C, Puigdollers P (2017) Rice blast forecasting models and their practical value: a review. Phytopathol Mediterr 56(2):187–216

5. Kaundal R, Kapoor AS, Raghava GP (2006) Machine learning techniques in disease forecasting: a case study on rice blast prediction. *BMC Bioinf* 7(1):485
6. Shafaullah MAK, Khan NA, Mahmood Y (2011) Effect of epidemiological factors on the incidence of paddy blast (*Pyricularia oryzae*) disease. *Pak J Phytopathol* 23(2):108–111
7. Rini P, Dipankar M, Naik BS (2017) Effect of different meteorological parameters on the development and progression of rice leaf blast disease in western Odisha. *Int J Plant Prot* 10(1):52–57
8. Bregaglio S, Titone P, Hossard L, Mongiano G, Savoini G, Piatti FM, Tamborini L (2017) Effects of agro-pedo-meteorological conditions on dynamics of temperate rice blast epidemics and associated yield and milling losses. *Field Crops Res* 212:11–22
9. Walker SH, Duncan DB Estimation of the probability of an event as a function of several
10. Wiki.fast.ai. (2010) Logistic regression—deep learning course Wiki [Online]. Available: [http://wiki.fast.ai/index.php/Logistic\\_Regression#Gradient\\_Descent](http://wiki.fast.ai/index.php/Logistic_Regression#Gradient_Descent)

# Assistive Technology for Students with Visual Impairments: A Resource for Teachers, Parents, and Students



Amit Sadh

**Abstract** Today is era of technology. We see everywhere technology. Technology is the skills, methods, and processes used to achieve goals. In this manner, there are special types of technology tools that can help people who learn understand and working differently. These specific tools are known as assistive technology (AT). Actually, AT is any software, device, or equipment which helps people with disability and works around there to make easier their life. Like in regard to visual impairment, Braille watch is an AT who helps the visual impaired person to know time. So, we can say that AT is one of essential parts of visually impaired child's academic or whole life. It allows VI to use their skills engaged in school environment and uses these technologies without helping other. The Rights of Person With Disability Act, 2016 (RPWD) mandated for assistive technology is one of crucial parts of visually impaired student. There are several researches which found that use of technology (AT) is lesser in rural area as compared to urban area in India. The scope of AT is fully determined by the knowledge of teachers and their knowledge of technology. The main theme of this paper is to potency of AT for visually impaired person. Here, in this scenario, some of special provisions have been also discussed in this paper. Its focuses on various types of technology and their use in life of visually impaired child.

**Keywords** AT · Schools · Technology · Braille · Visual impairments · Disability · (Person with disability (PWD) · Visual impairment (VI)

---

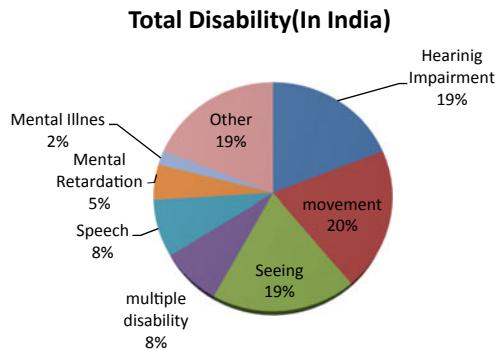
A. Sadh (✉)

Department of Secondary Education, Bikaner, Rajasthan, India  
e-mail: [amitsadh9898@gmail.com](mailto:amitsadh9898@gmail.com)

## 1 Introduction

As per census 2011, main data is as following

1. Total population: 121 Cr [1]
2. 2.68 Cr disabled (2.21% of population) [1]
3. In total, rural area 69% and 31% in urban area (disability static in India is as following in diagram) [1].



There are 5,033,431 visually impaired in India which is the world's largest disability population. Approximately, 285 million people are visually impaired worldwide of which thirty-nine million are visually impaired and 246 have low vision, according to WHO statistics.

Visual impairment (as per RPWD 2016) [2]

- (a) Blindness: “Blindness” means a condition where a person has any of the following conditions, after best correction—
  - (i) total absence of sight; or
  - (ii) visual acuity less than 3/60 or less than 10/200 (Snellen) in the better eye with best possible correction; or
  - (iii) limitation of the field of vision subtending an angle of less than 10°.

## 2 National Policy for Persons with Disabilities-(2006)

This policy implies that person with disability is important assets for India and society and also seeks to create a healthy environment that provides inclusion-opportunities-freedom and full inclusion in society [3].

The focus of this policy is

1. Disabilities prevention
2. Rehabilitation measures

3. Woman and children with disabilities–
4. Barrier-free environment
5. Issue of disability certificates
6. Social security
7. Promotion of non-governmental organizations (NGOs)
8. Collection of regular information on persons with disabilities–
9. Research, sports, recreation, and cultural life
10. Development of assistive technology
11. Free education.

### 3 Use of Assistive Technologies

Children with VI (CVI) may attend a mainstream school with a resource base. This arrangement also gives the advantage of pupils being socially integrated in an ordinary school community. Collectively, CVI attending a resources base may not feel so isolated or special [4] In school with a resource unit technology, the extra support that pupils receive while in mainstream class group may not be very different from that which might be provided in any school.

Normally in all educational institutes, all students come with a number of queries [3] So that special child also come there (schools/special institutes) for finding solution of their problems. Like how technology will enhance their working capabilities and solve their barrier? And what sort of technology helps for getting education like others? Such types of queries are differing from man to man and as per disability and needs like visual perception things are not important for visually impaired child as compared to hearing impaired child, and audio device is not useful for hard of hearing person, but visually impaired person like it most. So, we can say that teacher and manufacturer have proper knowledge that what sort of special requirements is required for these people. Person with special needs according to their needs requires some conceptual-structural changes in technology like adding or removing tools in the conventional technology for more accessibility [5].

Any things or technology which is enhancing working capacity or ability is known as assistive technology. In other words, we can say that AT means any adaptive device or service product equipment that increases independence, achievement participation, learning capacity of person with disability. AT enables people to live confident, and to involve in civic life and the labor market life. AT helps person with disability to live healthy independent and dignified lives and civic life. AT helps person with disabilities to actively participate in academic life like read-write, etc.

In wide context, AT is useful for creating new abilities for person with special needs including physical, intellectual, cognitive, learning, and sensory disability. AT helps students with visual impairment who use these in their educational institute like in completing assignment, examination, and other curriculum and co-curriculum activity. There are a lot of assistive technologies like Braille embosser, Braille display, audio devices, touch control device, CCTV, speech synthesizer, smart phone, etc. ATs

are highly imperative for helping VI student to achieve own goals and aims in life and getting success. AT also helps visually impaired child for social interaction with sighted and other person.

## 4 Special Technologies and Devices

Classifying visual acuity of any person is a very tough work. Most of the visually impaired students need some special types of special technology for getting productive and effective learning during their studies. Low-vision students require large font materials, low-vision device like magnifiers, lenses, CCTV, and other technology for good approach. The students with VI who normally use a recorded tape material or Braille printed books these days with technology having many option, a lot of devices and technologies make them confident and independent like using a simple reading application in mobile [5]. These devices are very useful for visually impaired and low-vision students. In the area of computer technology, innovation of speech synthesizer also performs a crucial part in academy life of VI students [6].

## 5 Auditory-Based Technologies

Auditory-based technology is helpful for visually impaired child like recording lesson or book by using tape recording for later use or review. Most auditory AT for visually impaired students employs synthetic speech. These include talking computer interfaces, reading machine [7], use of software like NVDA, which convert text to voice, talking calculator, and voice recognition computer. For reading purpose, the use of pre-recorded study material and speech recognition may crucially important to save energy, effort, and time of visually impaired student. There still are no AT that uses both 3D sound and signification to augment the auditory environment for visually impaired student.

## 6 Braille Watch



Braille watch is also another part of dot watch which is designed for visually impaired person. Both are same, actually Braille watch is portable hand wear technology that is used by visually impaired child for knowing time. Its process is very simple, by touching dial pad and noticing the embossment of Braille dot or arrangement of dialer. Both digital and analog versions are available. The analog watch has a protected glass cover that can be easily opened when the child wants to know time [8]. Dialer are tight, so during embossment, there is no movement in dialer. In digital form, the Braille dot changes its position as time changes. As we know that visually impaired student use Braille script, so in digital watch, they can easily read time in watch.

## 7 Talking Clock

Talking clock is a device that presents the time in the form of sound [9]. It is a recorded or live human voice service usually accessed by telephone that gives correct time to user or visually impaired person. The first speaking clock was introduced in France on February 14, 1933 [9]. It may present time merely as sound like as a telephone-based time service or clock for visually impaired person.



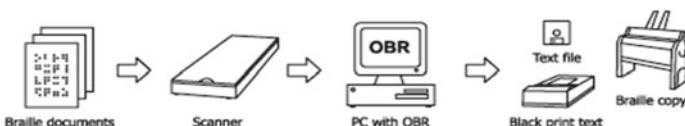
## 8 Optical Character Recognition (OCR)

OCR system provides person who is visually impaired with ability to scan printed text or font and then convert into audio format. The key element is OCR is scanning, recognition, and reading text. Now, there are many advancements in OCR technology. So, the current time OCR technology provides fine accuracy and amazing formatting capabilities like OrCam and MyEye. The OrCam is an OCR device that takes a picture of texts and relays the message to the user via a mini earpiece. Actually, OCR helps visually impaired students to scan and read printed material and convert into synthetic or digital speech. Using this technology, visually impaired student's education has become easier.

## 9 Braille Scanning Software

Optical Braille recognition (OBR) is Windows-based software that gives facility to the user to read Braille document on standard scanner either in single- or both-sided document. Its process is very simple, i.e.,

1. analyzes the Braille dot;
2. translates into normal text;
3. shows in computer screen.



## 10 Audio Devices

Mostly VI student prefer to use audio material in studies along with Braille. Visually impaired student normally uses recording material like cassette and recording machine for a lot of purposes, like recording their books, notes, and other study materials. They feel easy to submit their assignment in audio format as compared to Braille format. There are a lot of audio-format books, and software is available in online market like simply reading application, DAISY player, etc. Talking books are available in different format which run on prior or sophisticated audio devices.

Nowadays, digital accessible information system (DAISY) player plays a great role in education of VI child due to an audio advancement of its technology. The DAISY is technical standard for periodicals computerized and digital audio book, and it is the world's leading technology for digital audio talking books specially designed for visually impaired person. There are three types of DAISY books, i.e., as following [10]:

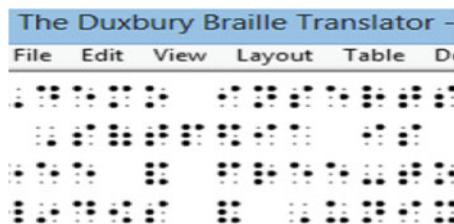
1. Audio-only DAISY, is the very usual technology. Its framework provides low data and a recorded data that the user listens when the book is played.
2. Text-only DAISY is different from audio-only because there are replacements of audio by text of books itself. It is used with the help of Braille display or text-to-speech (TTS). Bookshare.org is creating these DAISY books (text only).
3. The Cadillac in DAISY books is found in the full-text-audio DAISY book. This type of DAISY data provided both audio and text facility with synchronization. With the help of this advanced technology, user can easily hear TTS voice at will to judge grammatical information like spelling, etc. These types of books also run on player which directly not support only text format. Also other advanced features are involved in this DAISY player [10].



Daisy Player

## 11 Braille Translators

Braille displays and translators are especially designed for visually impaired person who normally uses Braille method. Reading and writing Braille with this technology is very simple method. It is a very simple process to reading and writing Braille using this software technology. Nowadays, the famous or general technology has furnished with Braille display.



### 11.1 Braille Embosser

It is a special printer that transfer text-data as Braille, or we can say that it is a device that can generate printed material using the Braille writing system for visually impaired person. By the use of specially designed software for translation of Braille script, a Braille document can be converted very easily, making Braille printing approachable and proficient. Nowadays, thousand of reading materials of different area and language of the world are embossed in Braille. With this, printed materials are easily available for visually impaired in very cheaper rate. Today, there lots of types Braille embosser are easily available in market and used for producing magazines and books for visually impaired. Operating these technologies is very easy. So, it is a widely used technology in the field of visually impaired person's education.

Mini desktop embosser are usual and can be established in space or place for visually impaired person like university, libraries, and special person center, as well as being personally owned by visually impaired. Belgian-made NV Interposing 55 is the fastest industrial Braille embosser, and its output is up to 800 Braille characters/second.



## 12 Screen Reader

Screen reader is application software which helps visually impaired person to use computer without help of sighted person. We can say that it is interface between operating system, its running application, and visually impaired person. Screen readers work closely with the operating system to provide information about menus, icons, folders, files, etc. There are two ways that this hardware can provide feedback to the user, i.e., speech and Braille. Screen reader software uses text-to-speech engine to translate on-screen data into speech which is heard by earphone, headphone, or speakers. There are several types of screen reader software used in different area of world, and in which, some are as following:

- NVDA
- Job Access With Speech (JAWS)
- VoiceOver (iOS)
- ZoomText fusion.



## 13 Braille Displays

It is an electro-mechanical device for showing Braille script sentences and characters. Actually, it is tactile device that consists of a row of special cells which are soft. It normally attached with keyboard which allows visually impaired to read data of display a single line at a single moment Braille script. When keyboard and Braille display are connected at same time, Braille displays make it possible for visually impaired child to use computer, read the display, and browsing Internet like sending and receiving email. Visually impaired users who cannot use monitor can easily use it to read text output. Deaf blind can also use this technology very effectively after proper training.

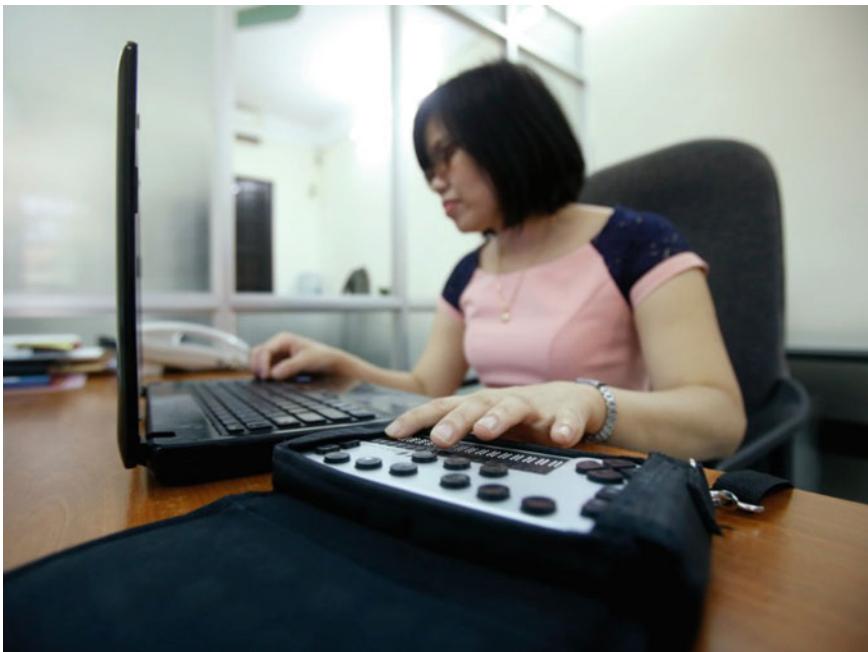


Diagram. A visually impaired woman uses a laptop and a Braille display

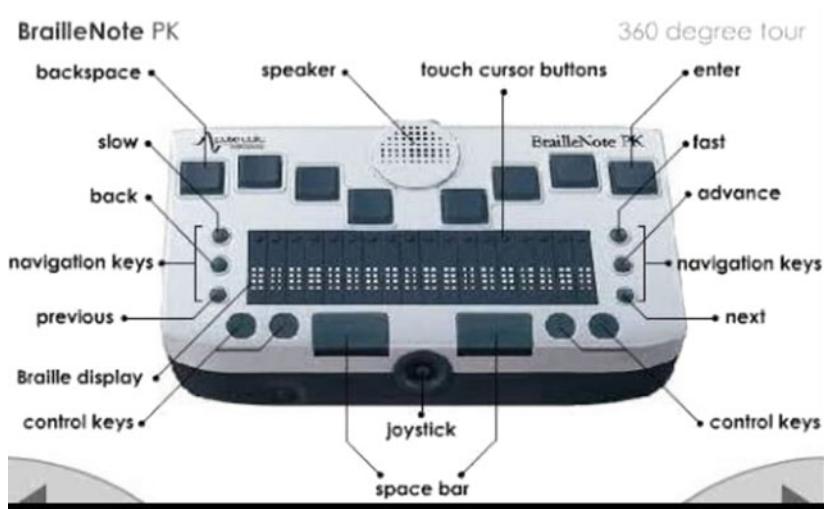
## 14 Google Talkback

Google Talkback is an accessibility service for Android users which help visually impaired person to interact with [11] mobile. It also uses text-to-speech engine. Talkback has a lot of features like spoken word, vibration, and other audio feedback

f or user to know what is going on screen and help user to interact with screen. Actually, this time it is the most frequently usable technology which is used by visually impaired person.

## 15 Braille Note-Takers

It is mini portable AT device especially made for visually impaired person. Braille note takers are used to take notes in Braille script. This device is a storing device with the use of Braille typewriter keyboard. The stored information is accessed by or Braille display or default speech synthesizer. An old note taker takes a lot of times to make education materials like notes. So, resultant of this is that visually impaired student could not take entire notes of session or class. But nowadays, it is not happening, because the technology of Braille note taking has become easy and fast. Braille note takers are helping to make notes in a faster way for visually impaired. The latest note taker device provides advance Web browsing, word processing, sending message, and other function.



## 16 Talking Calculator

A talking calculator has a built-in speech synthesizer that speaks loud on each operation like press number, symbol, etc. Talking calculators function like common calculators. This device is mostly used for visually impaired person. This can help the user to verify whether the operands have been entered correctly. It also speaks answer

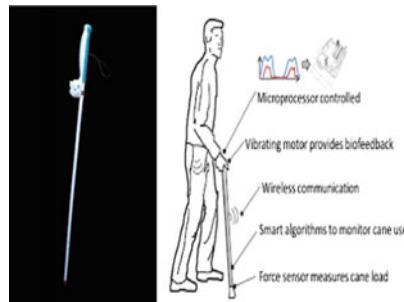
after finishing operation. There are a lot of variants which are available in market. A common feature is enlarging number so that low-vision person can also use this easily. This AT is available in market at cheaper price. The additional features are also added in talking calculator like alarm, music, and also reminder. Sometime, extra large number input button is also very useful for visually impaired child.



## 17 Smart Cane

Smart cane is an advanced cane used by visually impaired person for independent safety mobility, or we can say that smart cane is useful for visually impaired child for smart mobility. Normally, smart cane is white with folding variant. We can say that its transformation of white cane is frequently used by visually impaired person. White cane cannot detect overhanging object like open glass window, bunches of trees, or sign board. Then, the need for smart stick was felt [12].

These can solve many challenges and provide and guarantee visually impaired through safe mobility or independence. Visually impaired person can easily detect ground barrier, surface structure, etc., during his journey. This is not possible in white cane because white cane scratches a parking vehicle and impinges into other person.



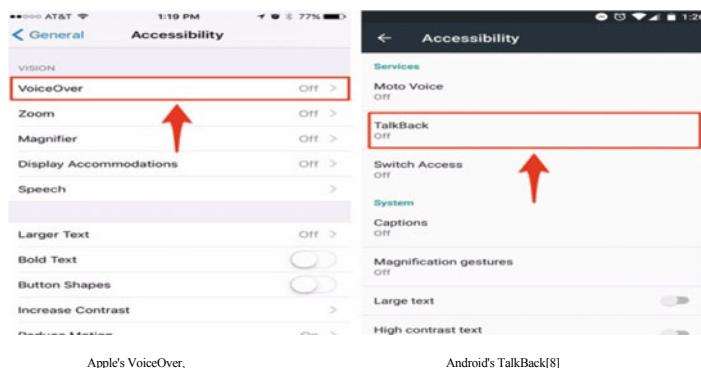
## 18 Smart Phone

When a sighted person uses a touch screen, smart phone, or tablet, he or she taps icons or slides a finger across the today's era smart phone which is not a unfamiliar term. Smart phone has become most important part of our life. While using smart

phone, user taps icons or slides finger across the display in order to make any event on screen. Nowadays, a famous or we can say that three top marketplace companies are Apple, Google, and Microsoft. Each includes both a screen magnifiers and a screen reader in their OS or devices [13].

Edit fields are areas that require the entry of information (such as the phone number you would like to dial or the text of an email you would like to send). When you double tap an edit field, the devices' on-screen keyboard appears. Most touch screen devices offer at least two ways to type.

- Standard typing: This is usually the default typing mode. Touch the screen until you find the character you wish to enter, or swipe your finger and listen as the new character is announced. When you find the key you want, perform a double tap. It is that simple [13].
- Touch typing: As you become familiar with the on-screen keyboard, you may wish to speed up your text and number entry. Touch typing mode allows you to find the key you want to enter by either touching the screen or sliding a finger across the on-screen keyboard until your device announces the character you want. At that point, simply lift your finger off the screen to enter and voice the character [13]. Then, you can return your finger to the screen to locate the next desired character, digit, or punctuation mark (See photo below [14]).



## 19 Conclusion

Visually impaired person's needs are greater than ever before. AT will continue to change the lives of visually impaired person as compared to previous decades. The increase of advance computing, mobile technologies, or other many electronics devices is expected to drive the field further toward the challenges and reality of creating usable AT.

Actually visually impaired person always struggle for education in periodic field episodic to their disability. As technology progresses day-by-day, resource for visually impaired person also increases very much. Today, we cannot imagine education of visually impaired student without using assistive technology. There are many online courses, and Web sites are available and specially designed for visually impaired person. Although educational bodies like school, college, and universities have been not as active to provide and ensure accessibility of learning materials and environment of this special person. But they can easily adopt some technology with the help of AT like audio-video technology etc [15]. If using AT in study, their effect will also be shown in education of student either it implement on normal child or special child. These technologies consist a variety of new things which consists a lot of types of devices' hardware and software applications which allow students who are visually impaired to get knowledge without any barrier.

## References

1. Census of India (2001) General population tables
2. Niemann S, Jacod N Helping children who are blind
3. <https://www.ncpedp.org/RPWdact2016>
4. Webster A, Roe J Children with visual impairment
5. Sharada R. P 1: Persons with visual impairments and their educational needs in India: use of special devices and assistive technologies
6. <https://slideplayer.com/slide/3540597/>
7. <https://www.brighthubeducation.com/special-ed-visual-impairments/74539-assistive-technology-for-students-with-visual-impairments/>
8. [https://en.wikipedia.org/wiki/Braille\\_watch](https://en.wikipedia.org/wiki/Braille_watch)
9. [https://en.wikipedia.org/wiki/Talking\\_clock](https://en.wikipedia.org/wiki/Talking_clock)
10. [http://www.mospi.gov.in/sites/default/files/reports\\_and\\_publication/statistical\\_publication/social\\_statistics/Chapter%208%20-National%20redressal.pdf](http://www.mospi.gov.in/sites/default/files/reports_and_publication/statistical_publication/social_statistics/Chapter%208%20-National%20redressal.pdf)
11. [https://en.wikipedia.org/wiki/Google\\_TalkBack](https://en.wikipedia.org/wiki/Google_TalkBack)
12. Siekierska E, Richard L, Louis B, Bill M, Peter P Enhancing spatial learning and mobility training of visually impaired people—a technical paper on the Internet-based tactile and audio-tactile mapping
13. <https://www.afb.org/blindness-and-low-vision/using-technology/cell-phones-tablets-mobile-touchscreen-smartphone>
14. <https://www.insider.com/how-blind-people-use-smartphones-2017-2>
15. [www.nfb.org](http://www.nfb.org)